

# BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email [info.bmjopen@bmj.com](mailto:info.bmjopen@bmj.com)

# BMJ Open

## Deriving and validating a risk prediction model for long COVID-19: protocol for an observational cohort study using linked Scottish data

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2021-059385
Article Type:	Protocol
Date Submitted by the Author:	23-Nov-2021
Complete List of Authors:	Daines, Luke; The University of Edinburgh College of Medicine and Veterinary Medicine, Usher Institute Mulholland, Rachel ; The University of Edinburgh College of Medicine and Veterinary Medicine, Usher Institute Vasileiou, Eleftheria; The University of Edinburgh, Usher Institute Hammersley, Vicky; The University of Edinburgh College of Medicine and Veterinary Medicine, Usher Institute Weatherill, David; The University of Edinburgh College of Medicine and Veterinary Medicine Katikireddi, Srinivasa; University of Glasgow, MRC/CSO Social & Public Health Sciences Unit Kerr, Steven; The University of Edinburgh College of Medicine and Veterinary Medicine, Usher Institute Moore, Emily; Public Health Scotland, Data Driven Innovation Pesenti, Elisa; The University of Edinburgh, Usher Institute Quint, Jennifer; Imperial College London, Respiratory Epidemiology, Occupational Medicine and Public Health Shah, Syed Ahmar; The University of Edinburgh Usher Institute of Population Health Sciences and Informatics Shi, Ting; The University of Edinburgh College of Medicine and Veterinary Medicine, Usher Institute Simpson, Colin; Victoria University of Wellington Robertson, Chris; University of Strathclyde, Department of Mathematics and Statistics Sheikh, Aziz; The University of Edinburgh College of Medicine and Veterinary Medicine, Usher Institute
Keywords:	PUBLIC HEALTH, COVID-19, Protocols & guidelines < HEALTH SERVICES ADMINISTRATION & MANAGEMENT

SCHOLARONE™  
Manuscripts

1  
2  
3  
4 **1 Deriving and validating a risk prediction model for long COVID-19:**  
5 **2 protocol for an observational cohort study using linked Scottish data**  
6  
7  
8

9 4 Luke Daines\*<sup>1</sup>, Rachel H Mulholland\*<sup>1</sup>, Eleftheria Vasileiou<sup>1</sup>, Vicky Hammersley<sup>1</sup>, David  
10 5 Weatherill<sup>1</sup>, Srinivasa Vittal Katikireddi<sup>2</sup>, Steven Kerr<sup>1</sup>, Emily Moore<sup>3</sup>, Elisa Pesenti<sup>4</sup>, Jennifer  
11 6 Quint<sup>5</sup>, Syed Ahmar Shah<sup>1</sup>, Ting Shi<sup>1</sup>, Colin R Simpson<sup>1,6</sup>, Chris Robertson<sup>3,7</sup>, Aziz Sheikh<sup>1</sup>  
12  
13  
14  
15

16 8 \*Contributed equally  
17  
18  
19

- 20  
21 10 1. Usher Institute, University of Edinburgh, Edinburgh, UK  
22 11 2. MRC/CSO Social & Public Health Sciences Unit, University of Glasgow, Glasgow, UK  
23 12 3. Public Health Scotland, Glasgow and Edinburgh, UK  
24 13 4. Institute of Cell Biology, University of Edinburgh, Edinburgh, UK  
25 14 5. Faculty of Medicine, National Heart & Lung Institute, Imperial College London, London,  
26 15 UK  
27 16 6. School of Health, Wellington Faculty of Health, Victoria University of Wellington,  
28 17 Wellington, NZ  
29 18 7. Department of Mathematics and Statistics, University of Strathclyde, Glasgow, UK  
30  
31  
32  
33  
34  
35  
36

37 20 **Corresponding author**  
38

39 21 Dr Luke Daines  
40

41 22 Usher Institute, University of Edinburgh,  
42

43 23 Doorway 3, Old Medical School, Teviot Place, Edinburgh United Kingdom  
44

45 24 Email: luke.daines@ed.ac.uk  
46  
47  
48

49 26 **Word count:** 3842 / 4000  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

# 1 **ABSTRACT**

## 2 **Introduction**

3 Coronavirus disease 2019 (COVID-19) is commonly experienced as an acute illness, yet  
4 some people continue to have symptoms that persist for weeks, or months (commonly  
5 referred to as “long-COVID”). It remains unclear which patients are at highest risk of  
6 developing long-COVID. In this protocol, we describe plans to develop a prediction model to  
7 identify individuals at risk of developing long-COVID.

## 8 **Methods and analysis**

9 We will use the national Early Pandemic Evaluation and Enhanced Surveillance of COVID-  
10 19 (EAVE II) platform, a population-level linked dataset of routine electronic healthcare data  
11 from 5.4 million individuals in Scotland. We will identify potential indicators for long-COVID  
12 by identifying patterns in primary care data linked to information from out-of-hours GP  
13 encounters, accident and emergency visits, hospital admissions, outpatient visits, medication  
14 prescribing/dispensing, and mortality. We will investigate the potential indicators of long-  
15 COVID by performing a matched analysis between those with a positive reverse  
16 transcriptase polymerase chain reaction (RT-PCR) test for Severe Acute Respiratory  
17 Syndrome 2 coronavirus (SARS-CoV-2) infection and two control groups; 1) individuals with  
18 at least one negative RT-PCR test and never tested positive; 2) the general population  
19 (everyone who did not test positive) of Scotland. Cluster analysis will then be used to  
20 determine the final definition of the outcome measure for long-COVID. We will then derive,  
21 internally and externally validate a prediction model to identify the epidemiological risk  
22 factors associated with long-COVID.

## 23 **Ethics and dissemination**

24 The EAVE II study has obtained approvals from the Research Ethics Committee (reference:  
25 12/SS/0201), and the Public Benefit and Privacy Panel for Health and Social Care  
26 (reference: 1920-0279). Study findings will be published in peer-reviewed journals and  
27 presented at conferences. Understanding the predictors for long-COVID and identifying the  
28 patient groups at greatest risk of persisting symptoms will inform future treatments and  
29 preventative strategies for long-COVID.

30  
31 **Word count:** 295/300

## 1 **Strengths and limitations of this study**

- 2 • We will use national data on ~99% of the Scottish population using the Early  
3 Pandemic Evaluation and Enhanced Surveillance of COVID-19 (EAVE II) platform.
- 4 • Our study will be unable to identify long-COVID patients who have not been in  
5 contact with healthcare services in Scotland.
- 6 • Identifying long-COVID using routinely collected electronic health records may be  
7 challenging due to the lack of a standardised definition and variation in coding  
8 practices across healthcare systems.
- 9 • To improve the identification of long-COVID (and associated clinical features) we  
10 intend to use free text in addition to the coded data available in electronic health  
11 records.
- 12 • We are actively involving individuals who have experienced long-COVID to shape the  
13 research and ensure relevance to patients and the public.

## 1 INTRODUCTION

2 In December 2019, an outbreak of a novel coronavirus was reported in Wuhan, China. The  
3 World Health Organization (WHO) declared the outbreak a global pandemic named  
4 coronavirus disease 2019 (COVID-19) caused by the Severe Acute Respiratory Syndrome 2  
5 (SARS-CoV-2) coronavirus. By November 2021, the WHO has reported over 240 million  
6 confirmed cases and at least five million deaths worldwide.[3] In the UK, more than nine  
7 million confirmed cases and over 140,000 deaths have been reported.[4]

8 The severity and duration of the acute SARS-CoV-2 infection varies widely. Most people are  
9 asymptomatic or experience mild-to-moderate symptoms, while a smaller proportion (10-  
10 15%) of cases experience more severe illness.[4] The majority of people recover after two to  
11 six weeks depending on disease severity.[5] However, some individuals have symptoms that  
12 last or recur for weeks or months after the initial acute infection.[5-20] Long-term effects of  
13 COVID-19 can present with a wide range of clinical features, relating to cardiovascular,  
14 neurological, respiratory and other organ systems, including mental health.[5-20] Common  
15 symptoms include fatigue, breathlessness, headaches, muscle weakness, joint pain and loss  
16 of taste or smell.[5,9,10,17-20]

17 Unified guidance to manage the long-term effects of COVID-19 in the United Kingdom (UK)  
18 has been developed by the National Institute for Health and Care Excellence (NICE),  
19 Scottish Intercollegiate Guidelines Network (SIGN) and the Royal College of General  
20 Practitioners (RCGP).[18] The guidance described two working case definitions of ongoing  
21 symptomatic COVID-19 (individuals with signs and symptoms of COVID-19 from four weeks  
22 to 12 weeks) and post-COVID-19 syndrome (individuals with signs and symptoms that  
23 develop during or following an infection consistent with COVID-19, continue for more than 12  
24 weeks and are not explained by an alternative diagnosis).[18] The term 'long-COVID'  
25 therefore commonly refers to those who continue to present signs and symptoms four weeks  
26 after acute COVID-19 infection i.e. both ongoing symptomatic COVID-19 and post-COVID-  
27 19 syndrome.[18]

28 In Scotland, patients with symptoms suggestive of long-COVID are advised to seek medical  
29 care from their general practitioner (GP).[21] Diagnostic codes (Read codes, version 2)  
30 within the Scottish GP electronic system were introduced in March 2021 using NICE-led  
31 working definitions of long-COVID.[22] Equivalent diagnostic codes were also introduced in  
32 Scotland's Scottish Clinical Coding Standards using International Classification of Diseases  
33 10<sup>th</sup> Revision (ICD-10) codes within secondary care data in February 2021.[23] The long-  
34 COVID diagnostic codes are available in the supplementary material.

1  
2  
3 1 Despite the progress in diagnostic coding, the prevalence and risk factors associated with  
4 2 long-COVID remain poorly understood, reflecting the lack of an agreed operational definition,  
5 3 the absence of diagnostic tests and the considerable variation in presentation. Reviews on  
6 4 the long-COVID literature have found that it was difficult to estimate the prevalence of  
7 5 persistent COVID-19 symptoms with certainty.[19,20] Therefore, alternative methods need to  
8 6 be adopted to identify those with long-COVID, so that the long-term consequences of  
9 7 COVID-19 illness can be better understood and individuals at highest risk of developing  
10 8 long-COVID can be identified early.[5,24-28] In this study, we aim to derive and validate a  
11 9 risk prediction model to estimate the probability that an individual will develop long-COVID.  
12 10 Our objectives are to: i) create an operational definition of long-COVID through studying  
13 11 health system interactions using a national linked healthcare dataset; ii) derive and validate  
14 12 a risk prediction model to estimate the probability of developing long-COVID; and iii)  
15 13 enhance the risk prediction model using machine learning.

## 14 **METHODS AND ANALYSIS**

### 15 **Study design and population**

16 We will undertake a national prospective population-based cohort study using the national  
17 Early Pandemic Evaluation and Enhanced Surveillance of COVID-19 (EAVE II) platform.[1,2]  
18 EAVE II comprises of routinely collected primary care, secondary care, laboratory and  
19 serology data from 5.4 million Scottish residents registered with a GP (~99% of the Scottish  
20 population) from February 2020.[1,2] We will primarily focus on adults (aged  $\geq 18$  years) and  
21 will follow-up this cohort until February 2023. We will consider extending the cohort to  
22 include children (aged  $<18$  years) if there are sufficient numbers of individuals in this age  
23 group.

### 24 **Inclusion/exclusion criteria**

25 Since the baseline population for this study is everyone registered with a GP, those who are  
26 not registered with a GP in Scotland will be excluded from the analyses.

### 27 **Sample size calculation**

28 We are using the whole population of Scotland and therefore sample size calculations are not  
29 applicable.

### 30 **Databases**

31 The EAVE II platform links a wide range of routine healthcare datasets using  
32 pseudonymised identifiers of National Health Service (NHS) Scotland's Community  
33 Healthcare Index (CHI). We will use these routinely collected data sources (described below)

1  
2  
3 1 to identify individuals with long-COVID and to determine their characteristics in the EAVE II  
4 2 cohort.

### 5 3 Primary care data

6  
7 4 Primary care data will be extracted from GP practices via EAVE II's trusted third party  
8 Albasoft Ltd.[1,2] GPs in the UK provide healthcare services that are free at the point of  
9 service and usually act as the first point of contact into the healthcare system. This data  
10 source captures all clinical and administrative activity at GPs and the characteristics of  
11 registered patients. These data are stored either as: 1) clinical codes; or 2) written free  
12 text.[27] The latter is used to capture detailed information on any encounter and may provide  
13 additional information not available in coded data. In order to include data from primary care  
14 encounters when GP practices are closed, we will use out-of-hours (OOH) records derived  
15 from the Public Health Scotland (PHS) Primary Care OOH Data Mart.[2]

### 16 13 Secondary care data

17 14 Activity in hospital-based care will be extracted from the Scottish Morbidity Record (SMR) 01  
18 which holds detailed information on hospital admissions, such as the specific area of clinical  
19 activity (specialty), the facility of care, patient management and new diagnoses.[2]  
20 Diagnoses in SMR01 will be extracted using ICD-10 codes.[28] For data on intensive care,  
21 we will use the Scottish Intensive Care Society Audit Group (SICSAG) dataset of all adult  
22 patients admitted to Intensive Care Units (ICU) and High Dependency Units (HDU) in  
23 Scotland.[2] For outpatient care, we will use the SMR00 dataset, which captures outpatient  
24 activity in specialist clinics such as physiotherapy.[2]

### 25 22 Laboratory data

26 23 All COVID-19 testing will be obtained from the Electronic Communication of Surveillance in  
27 Scotland (ECOSS) dataset. This surveillance data contains all reverse transcriptase  
28 polymerase chain reaction PCR (RT-PCR) tests carried out in Scotland.[2]

### 29 26 Vaccination data

30 27 COVID-19 vaccination data will be available from two sources: GP records and the Turas  
31 Vaccination Management Tool (TVMT), a web-based tool used to record community  
32 vaccinations in Scotland.[29]

### 33 30 Telehealth data

34 31 Telehealth in Scotland is operated by NHS 24 Scotland, which delivers telephone and online  
35 services [30]. We are specifically interested in the NHS 24 111 teleservice, which provides  
36 OOH advice. During the pandemic, this service was expanded to include a COVID-19



1  
2  
3 1 helpline which was used to provide advice and triage patients to COVID-19 Assessment  
4 2 Centres.[30]

5  
6  
7 3 Prescribing data

8  
9 4 Prescription data relating to all medications prescribed and dispensed in the community in  
10 5 Scotland will be extracted from the Prescribing Information System (PIS).[2] These  
11 6 medications are coded using the British National Formulary (BNF) code lists.[31] For  
12 7 medication data within hospitals, Hospital Electronic Prescribing and Medicines  
13 8 Administration (HEPMA) which are available for five Health Boards will be used.[2]

14  
15  
16  
17 9 Mortality data

18  
19 10 Mortality data will be taken from death registry data within the National Records of Scotland.  
20 11 These records hold information included on the death certificate, including cause(s) of death  
21 12 which are recoded using ICD-10 codes.[2]

22  
23  
24 13 Other data

25  
26 14 We will explore the use of other linkages available within the EAVE II platform. These  
27 15 include Scotland's Census 2011 from NHS Research Scotland (NRS) for information on  
28 16 ethnicity, disability, and occupation as part of the EAVE II sub-study for ethnic and social  
29 17 inequalities in COVID-19 outcomes in Scotland.[32] We will also consider linkages and  
30 18 comparisons to Generation Scotland's CovidLife surveys which launched in April 2020 to  
31 19 capture how COVID-19 has been affecting volunteers in the UK.[33]

## 20 **Determining an operational definition for long-COVID**

21 We will base our operational definition on the case definitions for the effects of COVID-19  
22 illness at different time periods developed by NICE[18]:

- 23 1. Acute COVID-19 infection: individuals with signs and symptoms of COVID-19 for up  
24 to 4 weeks
- 25 2. Ongoing symptomatic COVID-19: individuals with signs and symptoms of COVID-19  
26 from 4-12 weeks
- 27 3. Post-COVID-19 syndrome: individuals with signs and symptoms that develop during  
28 or following an infection consistent with COVID-19, continue for more than 12 weeks  
29 and are not explained by an alternative diagnosis. The post-COVID-19 syndrome  
30 usually presents with clusters of symptoms, often overlapping, which can fluctuate  
31 and change over time and can affect any organ system.

32 Long-COVID commonly refers to those who continue to present with signs and symptoms  
33 four or more weeks after acute COVID-19 infection, therefore our primary outcome will

1  
2  
3 1 include both ongoing symptomatic COVID-19 and post-COVID-19 syndrome. Our secondary  
4 2 outcome will focus on the clinical encounters suggestive of the post-COVID-19 syndrome.  
5 3 Further details are in the statistical analyses.  
6  
7

#### 8 4 **Population characteristics**

9  
10 5 Population characteristics will be explored to assess the risk factors for developing long-  
11 6 COVID and to account for any confounding in our analyses.  
12  
13

##### 14 7 Socio-demographics

15  
16 8 Age will be determined based on the available GP data and will be available as a continuous  
17 9 and categorical variable. Those aged over 100 will be truncated into the one group to  
18 10 overcome low sample size issues. Sex at birth will be included as a binary variable  
19 11 (female/male). Deprivation status will be derived from the Scottish Index of Multiple  
20 12 Deprivation (SIMD) 2020 quintile of the resident's postcode associated with their GP  
21 13 registration. Ethnicity data will also be included if completeness and quality of data is  
22 14 adequate. We will also consider other available information such as Body Mass Index (BMI)  
23 15 and smoking status (smoker, ex-smoker, non-smoker and unknown).  
24  
25  
26  
27  
28

##### 29 16 Geographical

30  
31 17 Area of residence in terms of NHS Scotland Health Boards and local authorities will be  
32 18 considered. Settlement type will be determined by the urban/rural 6-fold classification (UR6).  
33 19 Type of residence will also be considered such as private residence, care home and  
34 20 social/council housing if data are available.  
35  
36  
37

##### 38 21 Clinical characteristics

39  
40 22 Using diagnostic codes from the QCOVID algorithm,[34] we will identify the following  
41 23 conditions: a) cardiovascular; b) diabetes (type 1 and type 2); c) respiratory d) cancer (blood  
42 24 cancer, chemotherapy, lung or oral cancer, marrow transplant, radiotherapy); e)  
43 25 neurological; f) other conditions, such as liver cirrhosis, osteoporotic fracture, rheumatoid  
44 26 arthritis, systemic lupus erythematosus, sickle cell disease, venous thromboembolism, solid  
45 27 organ transplant, renal failure (chronic kidney disease stages 3-5 with or without dialysis or  
46 28 transplant).[34]  
47  
48  
49  
50

##### 51 29 Severity of acute COVID-19 illness

52  
53 30 Admission to hospital, any requirement for treatment in the ICU, and death will be used to  
54 31 categorise the severity of COVID-19 infection. We will define a COVID-19 hospitalisation as  
55 32 a RT-PCR confirmed positive test for SARS-CoV-2 in the 28 days prior to admission, or  
56 33 admission with an ICD-10 code for COVID-19. A COVID-19 ICU admission will be defined as  
57  
58  
59  
60

1  
2  
3 1 a RT-PCR confirmed positive test for SARS-CoV-2 in the 28 days prior to ICU admission. A  
4 2 COVID-19 death will be defined as dying within 28 days of confirmed or probable COVID-19.

### 3 **Missing data**

4 The amount of missing data will be examined for each variable of interest. Continuous  
5 variables, for example BMI, will be imputed using predictive mean matching or imputation by  
6 chained equations if appropriate. Categorical variables with missing data will have a distinct  
7 group of 'Unknown'. We will consider dealing with these missing categorical variables by  
8 either keeping the distinct group, imputing them using chained equations or removing them  
9 from the analysis. The latter will be a complete case analysis, which will reduce the total  
10 sample size.

### 11 **Statistical analyses**

#### 12 Developing an operational definition of long-COVID

13 We will firstly derive an operational definition for long-COVID by identifying patterns in  
14 clinical interactions within NHS Scotland services that may suggest long-COVID. A visual  
15 illustration of our intended methods is shown in Figure 1.

#### 16 *Indicators of long-COVID*

17 We are interested in a) GP interactions; b) hospital admissions; c) outpatient attendances; d)  
18 A&E visits; e) OOH encounters; f) NHS 24 telehealth interactions; g) medications (from GP  
19 prescribing and primary care pharmacy dispensing data); and h) all-cause mortality. Our  
20 primary focus will be on the GP data, with other healthcare data providing corroborative  
21 information (Step 1, Figure 1). Information within these electronic health datasets will serve  
22 as an investigative list of potential indicators for long-COVID.

23 For the healthcare services (sources a to f), we will investigate the frequency of interactions  
24 and the reasons for each interaction which will include any new diagnoses (categorised by  
25 body system), treatments, tests or procedures related to long-COVID. For medications (g),  
26 we will investigate the frequency and type of new prescriptions using British National  
27 Formulary (BNF) chapters. For all-cause mortality (h) we will record the causes of death  
28 using ICD-10 codes. Figure 2 summarises the different data sources and potential indicators  
29 of long-COVID we intend to investigate. The dataset will comprise of categorical binary  
30 variables (e.g., diagnosis or not) and numerical variables (e.g., number of consultations).

31 GP records provide a rich source of primary care data, with the coded data providing  
32 additional context to interactions such as the type of interaction (e.g., consultation,  
33 encounter, remote or face-to-face), referrals to specialty care and sick notes. For more  
34 detailed information on signs and symptoms that are indicative of long-COVID, we intend to

1 use written free text available from GP records. We will use natural language processing  
2 (NLP) to identify key words or phrases of signs and symptoms relating to long-COVID. This  
3 NLP model will be applied to all written free text using Computer-Assisted (diagnostic)  
4 Coding (CAC). This will create derived codes associated with the key words and phrases (1  
5 if the text mentions word or phrase, 0 otherwise).[35,36] These derived codes will be treated  
6 in a similar way to GP codes as discussed above. Figure 3 demonstrates this process of  
7 transforming the written GP free text into derived codes.

8 To initially explore potential long-COVID indicators, we will obtain summary level counts of  
9 all codes of interest within these healthcare datasets on individuals who tested positive and  
10 negative for COVID-19 using a RT-PCR test. We will count the frequency of these data  $\geq 4$   
11 weeks after the date of the test. This will inform which codes relating to potential long-COVID  
12 indicators will be extracted on a patient level for further analysis.

### 13 *Matched analysis.*

14 To identify which of these long-COVID indicators are most important, we will perform a  
15 matched analysis (Step 2, Figure 1). The exposed group will be defined as the first date an  
16 individual tested positive for COVID-19 using a RT-PCR test. Two control groups will be  
17 assigned: 1) individuals who have had at least one negative RT-PCR test and have never  
18 tested positive up to the date of the exposed match testing positive; and 2) the general  
19 population (everyone who did not test positive) of Scotland. These control groups will be  
20 investigated in turn.

21 We will use risk-set matching in a 3:1 ratio by time-varying propensity score matching. This  
22 will be based on the likelihood of testing positive for COVID-19 and will consider  
23 incorporating the following characteristics: sex, age, geography, comorbidities, risk factors,  
24 number of previous SARS-CoV-2 tests, deprivation status and urban-rural settlement. The  
25 adequacy of the matching will be assessed by checking for imbalance of the individual  
26 covariates across exposure groups.

27 For the matched analysis, each potential long-COVID indicator will be treated as its own  
28 dependent variable in turn. Follow-up will begin from four weeks after the exposed tested  
29 positive for COVID-19. Follow-up will end on either the date of event (if indicator is a binary  
30 variable), the control testing positive for COVID-19, death from any cause or the end of the  
31 follow-up period. Controls who have a positive test will be eligible to be included in the  
32 exposed group. Since this is a live cohort, we will update the end of follow-up on biannual to  
33 3-month basis until February 2023.

34 The long-COVID indicators will be compared between the exposed and control groups using  
35 statistical tests such as two-sample proportions test (for binary indicators), two-sample t-

1 tests (for continuous indicators), Kaplan-Meier curves to inspect cumulative incidence, and  
2 survival analysis to look at the potential impact of interventions on long-COVID symptoms.

3 We also plan to conduct similar analyses on the whole cohort without propensity score  
4 matching. We will consider stratifying by age and sex if numbers allow.

#### 5 *Cluster analysis*

6 Clusters of long-COVID presentations in the exposed group will be investigated further,  
7 using the long-COVID indicators as our clustering input (Step 3, Figure 1). The indicators will  
8 be summarised using a window of four or more weeks after initially testing positive (e.g., the  
9 number of interactions  $\geq 4$  weeks after the test). These indicators will not include the  
10 diagnostic codes for long-COVID since we are aiming to provide a more accurate alternative  
11 to this measurement of long-COVID.

12 We will explore both hierarchical clustering and k-means clustering, using distance  
13 measurements such as the Gower Distance which is a suitable measurement of similarity for  
14 mixed categorical and numeric data.[37] We will also investigate clusters based on latent  
15 class analysis. We will then internally validate these clusters using statistics such as the  
16 silhouette coefficient and the Dunn index.[38,39] Comparisons between the clusters and  
17 long-COVID diagnostic codes will be undertaken for validation. The final set of clusters of  
18 long-COVID indicators will serve as our operational definition for long-COVID.

#### 19 *Sensitivity analyses*

20 We will perform a variety of sensitivity analyses to test the robustness of our long-COVID  
21 definition. This includes evaluating the start of follow-up to 12 weeks, to explore whether the  
22 alternative outcome definition of 'post-COVID-19 syndrome' display different clinical  
23 pathways. We will also investigate the patterns in the long-COVID indicators associated with  
24 the diagnostic long-COVID codes. To capture those who may be suffering from long-COVID  
25 but did not formally test positive for COVID-19, we will investigate the long-COVID indicators  
26 in the general population. We will also stratify by time-period, for example during the different  
27 peaks of positive cases in Scotland (e.g., March 2020 to July 2020, August 2020 to April  
28 2021).[4] This will also reflect the dominant COVID-19 variants during the different waves of  
29 infection.

#### 30 *Deriving and validating a risk prediction model for long-COVID*

31 We will use the transparent reporting of a multivariable prediction model for individual  
32 prognosis or diagnosis (TRIPOD) guidelines to report the derivation and validation of the  
33 long-COVID prediction model (see completed checklist in the supplementary material).[40]

1  
2  
3 1 The model will be derived using data from everyone in the cohort (defined above) who  
4  
5 2 received a positive PCR test.

### 6 7 3 *Descriptive analysis*

8  
9 4 We will begin analysis by conducting descriptive analyses to visually inspect and summarise  
10  
11 5 the types of potential long-COVID presentations within the clusters. Next, summaries of the  
12  
13 6 geographical, sociodemographic and risk factor profile of those presenting with the long-  
14  
15 7 COVID clusters will be reported.

### 16 17 8 *Outcome*

18  
19 9 The outcome for the risk prediction model will be the derived operational definition of long-  
20  
21 10 COVID defined from the cluster analysis. This will be dependent on the number of optimum  
22  
23 11 clusters and the different classifications of long-COVID presentation. Depending on the  
24  
25 12 clusters, we will classify our outcome into a binary variable of belonging to one (or more)  
26  
27 13 cluster(s) (1) or otherwise (0).

### 28 29 14 *Predictor variables*

30  
31 15 Predictors for the risk prediction model will consist of the patient characteristics, including  
32  
33 16 information on socio-demographics, geographical, clinical comorbidities and severity of  
34  
35 17 COVID-19 infection. These will be a mixture of continuous, binary and categorical variables.  
36  
37 18 Continuous variables will be tested for linearity and for more flexible relationships (using  
38  
39 19 smooth splines). Groupings of continuous variables will be explored if necessary.

### 40 41 20 *Type of model*

42  
43 21 We intend to use a multivariable logistic regression model.

### 44 45 22 *Selection of predictors*

46  
47 23 We will build our model using stepwise selection based on the Akaike's Information Criterion  
48  
49 24 (AIC) and Bayesian information criterion (BIC). To assess the fit of the model parameters,  
50  
51 25 the maximum likelihood ratio test will also be used.

### 52 53 26 *Model evaluation/performance*

54  
55 27 To evaluate the model's goodness of fit, we will use appropriate performance evaluation  
56  
57 28 metrics such as the area under the ROC curve (captures the accuracy of the model  
58  
59 29 discriminating between the outcome), and the calibration plot and slope (visualises the  
60  
30 observed vs predicted values). Other evaluation measures such as the specificity (true  
31  
32 negative rate), sensitivity (true positive rate) and accuracy will be considered if appropriate.  
We will also directly compare the predicted and observed values.



### 1 *Model validation*

2 The model will be internally validated using k-fold cross validation. We will validate using  
3 different time periods as specified by one of the discussed sensitivity analyses in the  
4 clustering analysis. We will explore opportunities for external validation, comparison and  
5 meta-analysis with other long-COVID initiatives.

### 6 *Risk groups*

7 To categorise the output of the model for further use by clinicians and COVID-19 patients,  
8 we will consider stratifying patients into risk groups based on the predictive probabilities in  
9 the multivariable model, for example three groups of low, moderate and high risk of long-  
10 COVID.

### 11 *Sensitivity analysis*

12 Sensitivity analyses for the risk prediction algorithm will depend on the outcomes from the  
13 sensitivity analyses from Objective 1 of developing an operational definition for long-COVID.

### 14 *Enhancing the prediction model using machine learning*

15 Advances in machine learning will be utilised to enhance the development and validation of  
16 the prediction model. Specifically, we will systematically explore the use of supervised  
17 learning algorithms such as penalised models (e.g., LASSO regression), naïve Bayes  
18 classifier, gradient boosting decision trees and random forests to further improve the  
19 prediction model developed with traditional statistical methods. We will also consider using  
20 ensemble learning methods to strategically combine multiple models to obtain better  
21 predictive performance.

### 22 **Patient and public involvement**

23 Lay input has shaped the development of this research and will continue throughout the  
24 project through the patient and public involvement (PPI) co-applicant, the EAVE II Public  
25 Advisory Group (PAG), and long-COVID Scotland. PPI members will collaborate with the  
26 research team to provide real world perspectives when analysing and interpreting study  
27 findings ensuring that the work considers the needs, interests and concerns of patient and  
28 public members.

### 29 **ETHICS AND DISSEMINATION**

#### 30 *Ethical approval*

31 This study forms part of the EAVE II project which is investigating epidemiological risk  
32 factors of COVID-19 disease. EAVE II has already obtained permissions from the Research  
33 Ethics Committee (REC reference: 12/SS/0201), NHS Research and Development, GPs,

1 NHS Health Boards and the Public Benefit and Privacy Panel (PBPP) for Health and Social  
2 Care (reference number: 1920-0279).

### 3 Dissemination

4 To ensure the greatest impact of our findings, we will actively disseminate to three key  
5 audiences: policy/public health, academic and community-based.

#### 6 *Policy and Public Health*

7 Findings from this project will provide evidence to help NHS Scotland and other international  
8 policy makers identify groups of the population who are at most risk of long-COVID and  
9 related complications. We will work with partners in NHS Scotland, Public Health Scotland  
10 (PHS) and the Scottish Government to establish the best ways of disseminating these  
11 results and influencing policy. We also plan to disseminate findings through the National  
12 Core Studies programme, a UK government initiative supported by Health Data Research  
13 UK (HDRUK), in partnership with the Office for National Statistics (ONS).

#### 14 *Academic*

15 We will communicate our findings through presentations at major national and international  
16 scientific meetings and through publications in relevant peer-reviewed journals. We will  
17 publish our code and data dictionary on the EAVE II's GitHub repository  
18 (<https://github.com/EAVE-II>).

#### 19 *Community-based*

20 We will write lay summaries of publications and create infographics to further communicate  
21 our findings via press releases, public and patient engagement events, social media and the  
22 EAVE II website (<https://www.ed.ac.uk/usher/EAVE-ii>).

23

24



## 1 **AUTHORS CONTRIBUTIONS**

2 AS conceived this manuscript. RM and LD led the writing of the manuscript. Statistical  
3 methods were reviewed by EAVE II's lead statistician CR and the other co-authors VK, LD,  
4 SAS and SK. All authors reviewed the manuscript.

## 5 **FUNDING STATEMENT**

6 This work was supported by the Chief Scientist Office, grant number COV/LTE/20/15. EAVE  
7 II is supported by a grant (MC\_PC\_19075) from the Medical Research Council; a grant  
8 (MC\_PC\_19004) from BREATHE--The Health Data Research Hub for Respiratory Health,  
9 funded through the UK. Research and Innovation Industrial Strategy Challenge Fund and  
10 delivered through Health Data Research UK; a grant from the Data and Connectivity  
11 National Core Study, led by Health Data Research UK in partnership with the Office for  
12 National Statistics and funded by UK Research and Innovation (grant ref MC\_PC\_20058);  
13 Public Health Scotland; and the Scottish Government Director General for Health and Social  
14 Care. SVK acknowledges funding from a NRS Senior Clinical Fellowship (SCAF/15/02), the  
15 Medical Research Council (MC\_UU\_00022/2) and the Scottish Government Chief Scientist  
16 Office (SPHSU17).

## 17 **COMPETING INTERESTS STATEMENT**

18 AS is a member of the Scottish Government Chief Medical Officer's COVID-19 Advisory  
19 Group and its Standing Committee on Pandemics. He is a member of the UK Government's  
20 Risk Stratification Subgroup and Astra-Zeneca's Thrombotic Thrombocytopenic Taskforce.  
21 All roles are unremunerated. SVK was co-chair of the Scottish Government's Expert  
22 Reference Group on Ethnicity and COVID-19 and a member of the UK Government's  
23 Scientific Advisory Group on Emergencies (SAGE) subgroup on ethnicity. All other authors  
24 declare no competing interests.

## 25 **ACKNOWLEDGEMENTS**

26 We are grateful for the support from Public Health Scotland (PHS) and Albasoft Ltd on the  
27 development of the protocol and data extraction. Methods were also developed with support  
28 from the CONVALESCENCE study analysts. We would also like to acknowledge the PPI  
29 support from the EAVE II PAG members and long-COVID Scotland.

## 30 **DATA AVAILABILITY**

31 No additional data available.

## 1 REFERENCES

1. Simpson CR, Robertson C, Vasileiou E, et al. Early Pandemic Evaluation and Enhanced Surveillance of COVID-19 (EAVE II): protocol for an observational study using linked Scottish national data. *BMJ Open* 2020;10:e039097.
2. 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. DOI: <https://doi.org/10.1093/ije/dyab028>
3. World Health Organization. WHO Coronavirus Disease (COVID-19) Dashboard. Available from: <https://covid19.who.int/> (Accessed November 2021)
4. Public Health Scotland. COVID-19 in Scotland Daily Dashboard. Available from: [https://public.tableau.com/profile/phs.covid.19#!/vizhome/COVID-19DailyDashboard\\_15960160643010/Overview](https://public.tableau.com/profile/phs.covid.19#!/vizhome/COVID-19DailyDashboard_15960160643010/Overview) (Accessed November 2021)
5. World Health Organization. What we know about Long-term effects of COVID-19. Available from: [https://www.who.int/docs/default-source/coronaviruse/risk-comms-updates/update54\\_clinical\\_long\\_term\\_effects.pdf?sfvrsn=3e63eee5\\_8](https://www.who.int/docs/default-source/coronaviruse/risk-comms-updates/update54_clinical_long_term_effects.pdf?sfvrsn=3e63eee5_8) (Accessed November 2021)
6. Rayner C, Lokugame AU, Molokhia M. Covid-19: Prolonged and relapsing course of illness has implications for returning workers. *BMJ* 2020. Available from: <https://blogs.bmj.com/bmj/2020/06/23/covid-19-prolonged-and-relapsing-course-of-illness-has-implications-for-returning-workers/> (Accessed November 2021)
7. Carfi A, Bernabei R, Landi F. Persistent Symptoms in Patients After Acute COVID-19. *JAMA* 2020;324(6):603-605.
8. del Rio C, Collins LF, Malani P. Long-term Health Consequences of COVID-19. *JAMA* 2020;324(17):1723-4.
9. Huang C, Huang L, Wang Y, et al. 6-month consequences of COVID-19 in patients discharged from hospital: a cohort study. *Lancet* 2020;397(10270):220-32
10. Evans RA, McAuley H, Harrison EM, et al. Physical, cognitive, and mental health impacts of COVID-19 after hospitalisation (PHOSP-COVID): a UK multicentre, prospective cohort study. *Lancet Respir Med*. 2021. DOI: [https://doi.org/10.1016/S2213-2600\(21\)00383-0](https://doi.org/10.1016/S2213-2600(21)00383-0)
11. Ayoubkhani D, Khunti K, Nafilyan V, et al. Epidemiology of post-COVID syndrome following hospitalisation with coronavirus: a retrospective cohort study. <https://doi.org/10.1101/2021.01.15.21249885>
12. Mitrani RD, Dabas N, Goldberger JJ. COVID-19 cardiac injury: Implications for long-term surveillance and outcomes in survivors. *Heart Rhythm* 2020;17(11):1984-1990.
13. Mao L, Jin H, Wang M et al. Neurologic Manifestations of Hospitalized Patients With Coronavirus Disease 2019 in Wuhan, China. *JAMA Neurol* 2020;77(6):683–690.
14. George PM, Wells AU, Jenkins RG. Pulmonary fibrosis and COVID-19: the potential role for antifibrotic therapy. *Lancet Respir Med* 2020;8(8):807-815.
15. Rajkumar RP. COVID-19 and mental health: A review of the existing literature. *Asian J Psychiatr* 2020;52:102066.
16. ZOE app: One in 20 people likely to suffer from 'Long COVID', but who are they? Available from: <https://covid.joinzoe.com/post/long-covid> (Accessed November 2021)

17. Office for National Statistics, Prevalence of ongoing symptoms following coronavirus (COVID-19) infection in the UK: 4 November 2021. Available from:  
<https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/prevalenceofongoingsymptomsfollowingcoronaviruscovid19infectionintheuk/4november2021> (Accessed November 2021)
18. NICE guideline. COVID-19 rapid guideline: managing the long-term effects of COVID-19. Available from:  
<https://www.nice.org.uk/guidance/ng188> (Accessed November 2021)
19. NIHR: Living with Covid19 – Second review. Available from:  
<https://evidence.nihr.ac.uk/themedreview/living-with-covid19-second-review/> (Accessed November 2021)
20. Usher Network for COVID-19 Evidence Reviews. What is the prevalence of the post-COVID-19 syndrome? Available from:  
[https://www.ed.ac.uk/files/atoms/files/uncover\\_028-01\\_summary\\_prevalence\\_of\\_post\\_covid-19\\_syndrome\\_2april.pdf](https://www.ed.ac.uk/files/atoms/files/uncover_028-01_summary_prevalence_of_post_covid-19_syndrome_2april.pdf) (Accessed November 2021)
21. NHS Inform. Longer-term effects of COVID-19 (long COVID).  
<https://www.nhsinform.scot/illnesses-and-conditions/infections-and-poisoning/coronavirus-covid-19/coronavirus-covid-19-longer-term-effects-long-covid> (Accessed November 2021)
22. Scottish Government. Management and recording of the long-term effects of COVID-19 (Long COVID). March 2020. Available from:  
<https://www.scimp.scot.nhs.uk/wp-content/uploads/CMO-Letter-09.03.21.pdf> (Accessed November 2021)
23. Public Health Scotland. Scottish Clinical Coding Standards. Number 27.  
<https://www.isdscotland.org/products-and-services/terminology-services/clinical-coding-guidelines/Docs/Scottish-clinical-coding-standards-Feb-2021-No-27.pdf> (Accessed November 2021)
24. Yelin D, Wirtheim E, Vetter P, et al. Long-term consequences of COVID-19: research needs. *Lancet Infect Dis* 2020;20(10):1115-1117.
25. Marshall M. The lasting misery of coronavirus long-haulers. *Nature*. 2020;585(7825):339-341.
26. Rimmer A. Covid-19: Impact of long term symptoms will be profound, warns BMA. *BMJ* 2020;370:m3218.
27. Scottish Clinical Information Management in Practice (SCIMP). SCIMP Guide to Read Codes. Available from:  
<https://www.scimp.scot.nhs.uk/better-information/clinical-coding/scimp-guide-to-read-codes> (Accessed November 2021)
28. Information Services Division. SMR Datasets. General Clinical Information. Available from: <https://www.ndc.scot.nhs.uk/Data-Dictionary/SMR-Datasets/General-Clinical-Information/Diagnostic-Section/General%20Information%20on%20the%20ISCD.asp> (Accessed November 2021)
29. Turas. Turas Vaccination Management tool. Available from:  
<https://learn.nes.nhs.scot/42708/turas-vaccination-management-tool> (Accessed November 2021)
30. NHS24. Available from: <https://www.nhs24.scot/111/when-to-phone-111/> (Accessed November 2021)
31. National Institute for Health and Care Excellence (NICE). BNF. June 2021. Available from: <https://bnf.nice.org.uk/> (Accessed November 2021)
32. University of Glasgow. Exploring Inequalities in COVID-19 in Scotland. Available from:  
<https://www.gla.ac.uk/researchinstitutes/healthwellbeing/research/mrccso>

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60
- 1 [socialandpublichealthsciencesunit/aboutus/covid19/inequalitiescovid/](https://socialandpublichealthsciencesunit/aboutus/covid19/inequalitiescovid/)  
2 (Accessed November 2021)  
3  
4 33. Generation Scotland. CovidLife. Available from:  
5 <https://www.ed.ac.uk/generation-scotland/for-researchers/covidlife>  
6 (Accessed November 2021)  
7  
8 34. Clift AK, Coupland CA, Keogh RH, et al. Living risk prediction algorithm  
9 (QCOVID) for risk of hospital admission and mortality from coronavirus 19  
10 in adults: national derivation and validation cohort study. *BMJ*  
11 2020;371:m3731.  
12  
13 35. Nguyen AN, Truran D, Kemp M, et al. Computer-Assisted Diagnostic  
14 Coding: Effectiveness of an NLP-based approach using SNOMED CT to  
15 ICD-10 mappings. *AMIA Annu Symp Proc*. 2018:807-816.  
16  
17 36. Koleck TA, Dreisbach C, Bourne PE, et al. Natural language processing of  
18 symptoms documented in free-text narratives of electronic health records:  
19 a systematic review. *J Am Med Inform Assoc*. 2019;26(4):364-379  
20  
21 37. Gower JC. A General Coefficient of Similarity and Some of Its Properties.  
22 *Biometrics* 1971;1:857-871.  
23  
24 38. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and  
25 validation of cluster analysis. *J Comput Appl Math* 1987;20:53-65.  
26  
27 39. Dunn J.C. A Fuzzy Relative of the ISODATA Process and Its Use in  
28 Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*  
29 1973;3(3)32-57, DOI: 10.1080/01969727308546046  
30  
31 40. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a  
32 multivariable prediction model for Individual Prognosis or Diagnosis  
33 (TRIPOD): explanation and elaboration. *Annals of internal medicine*.  
34 2015;162(1):W1-73.  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Review only

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1 **FIGURE LEGENDS**

2 Figure 1: Schematic diagram of methods for developing an operational definition for Long  
3 COVID.

4 Figure 2: Data linkage diagram of long-COVID indicators and their data sources within the  
5 EAVE II cohort.

6 Figure 3: Schematic diagram of Natural Language Processing (NLP) and Computer Assisted  
7 Coding (CAC) on GP free text

8  
9  
10  
11  
12  
13

For peer review only

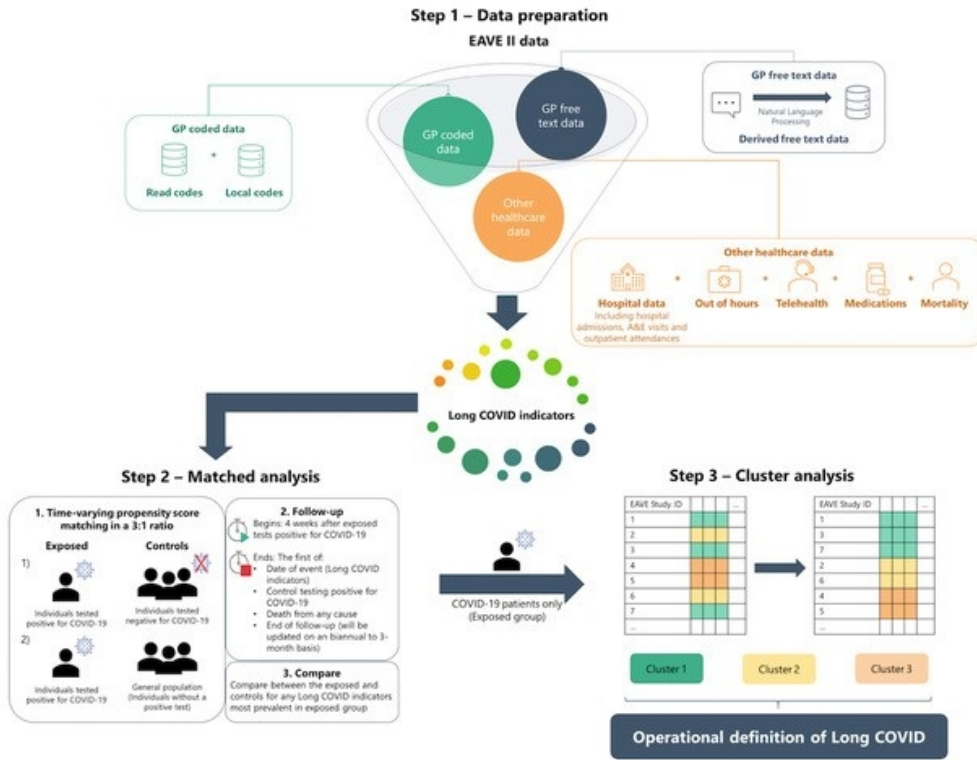
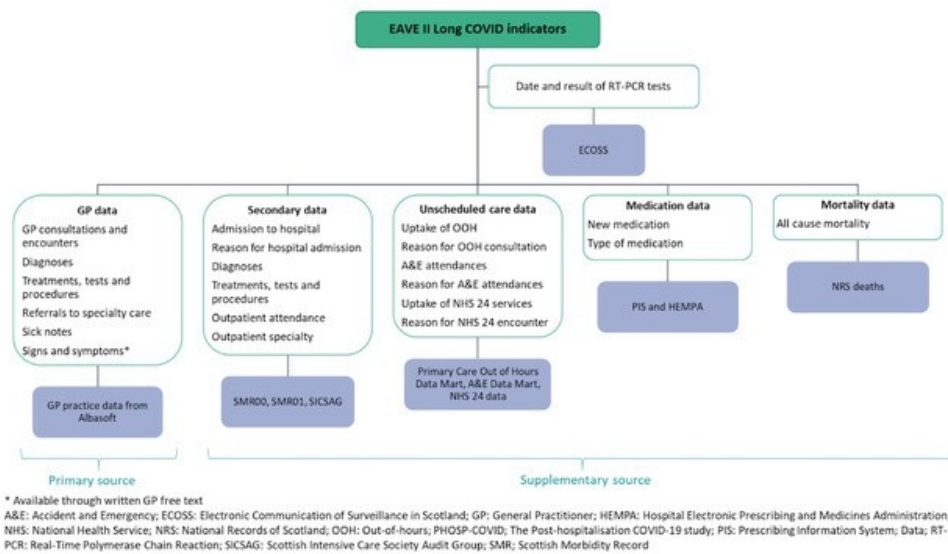


Figure 1: Schematic diagram of methods for developing an operational definition for Long COVID.

52x40mm (300 x 300 DPI)





26 Figure 2: Data linkage diagram of long-COVID indicators and their data sources within the EAVE II cohort.

27 54x32mm (300 x 300 DPI)

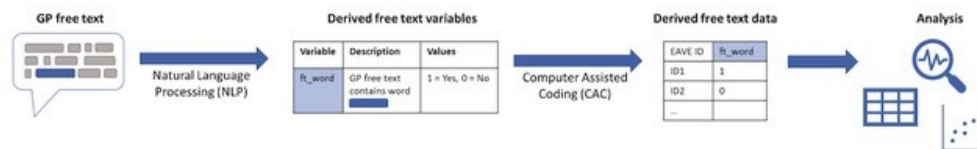


Figure 3: Schematic diagram of Natural Language Processing (NLP) and Computer Assisted Coding (CAC) on GP free text

54x9mm (300 x 300 DPI)



# Deriving and validating a risk prediction model for long COVID-19: protocol for an observational cohort study using linked Scottish data

## SUPPLEMENTARY FILE

Diagnostic codes for Long COVID

*Local codes introduced in Scotland<sup>1</sup>*

Clinical Computer System	Code	Description
EMIS PCS	^ESCT1348648	Ongoing symptomatic COVID-19
EMIS PCS	^ESCT1348645	Post-COVID-19 syndrome
Vision	A7955	Ongoing symptomatic COVID-19
Vision	AyuJC	Post-COVID-19 syndrome

*ICD-10 emergency use codes for conditions related to COVID-19<sup>2</sup>*

ICD-10 Code	Description
U07.3	Personal history of COVID-19
U07.4	Post-COVID-19 condition
U07.5	Multisystem inflammatory syndrome associated with COVID-19

Tripod checklist: Prediction model development and validation<sup>3</sup>

Section/Topic	Item	Checklist Item	Page
<b>Title and abstract</b>			
Title	1	D;V Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	D;V Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	2
<b>Introduction</b>			
Background and objectives	3a	D;V Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	4/5
	3b	D;V Specify the objectives, including whether the study describes the development or validation of the model or both.	5
<b>Methods</b>			
Source of data	4a	D;V Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	5/6
	4b	D;V Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	5/6
Participants	5a	D;V Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	5/6
	5b	D;V Describe eligibility criteria for participants.	5
	5c	D;V Give details of treatments received, if relevant.	NA
Outcome	6a	D;V Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	7
	6b	D;V Report any actions to blind assessment of the outcome to be predicted.	NA
Predictors	7a	D;V Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	8
	7b	D;V Report any actions to blind assessment of predictors for the outcome and other predictors.	NA
Sample size	8	D;V Explain how the study size was arrived at.	5
Missing data	9	D;V Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	9
Statistical analysis methods	10a	D Describe how predictors were handled in the analyses.	12
	10b	D Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	12
	10c	V For validation, describe how the predictions were calculated.	NA
	10d	D;V Specify all measures used to assess model performance and, if relevant, to compare multiple models.	NA
	10e	V Describe any model updating (e.g., recalibration) arising from the validation, if done.	NA
Risk groups	11	D;V Provide details on how risk groups were created, if done.	NA
Development vs. validation	12	V For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	NA
<b>Results</b>			
Participants	13a	D;V Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	NA
	13b	D;V Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	NA
	13c	V For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	NA
Model development	14a	D Specify the number of participants and outcome events in each analysis.	NA
	14b	D If done, report the unadjusted association between each candidate predictor and outcome.	NA
Model specification	15a	D Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	NA
	15b	D Explain how to use the prediction model.	NA
Model performance	16	D;V Report performance measures (with CIs) for the prediction model.	NA
Model-updating	17	V If done, report the results from any model updating (i.e., model specification, model performance).	NA
<b>Discussion</b>			
Limitations	18	D;V Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	NA
Interpretation	19a	V For validation, discuss the results with reference to performance in the development data, and any other validation data.	NA
	19b	D;V Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	NA
Implications	20	D;V Discuss the potential clinical use of the model and implications for future research.	NA
<b>Other information</b>			
Supplementary information	21	D;V Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	NA
Funding	22	D;V Give the source of funding and the role of the funders for the present study.	

## REFERENCES

- 1) Scottish Government. Management and recording of the long-term effects of COVID-19 (Long COVID). March 2020. Available from: <https://www.scimp.scot.nhs.uk/wp-content/uploads/CMO-Letter-09.03.21.pdf> (Accessed November 2021)
- 2) Public Health Scotland. Scottish Clinical Coding Standards. Number 27. <https://www.isdscotland.org/products-and-services/terminology-services/clinical-coding-guidelines/Docs/Scottish-clinical-coding-standards-Feb-2021-No-27.pdf> (Accessed November 2021)
- 3) Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine*. 2015;162(1):W1-73.

For peer review only

# BMJ Open

## Deriving and validating a risk prediction model for long COVID-19: protocol for an observational cohort study using linked Scottish data

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2021-059385.R1
Article Type:	Protocol
Date Submitted by the Author:	05-Apr-2022
Complete List of Authors:	Daines, Luke; The University of Edinburgh College of Medicine and Veterinary Medicine, Usher Institute Mulholland, Rachel ; The University of Edinburgh College of Medicine and Veterinary Medicine, Usher Institute Vasileiou, Eleftheria; The University of Edinburgh, Usher Institute Hammersley, Vicky; The University of Edinburgh College of Medicine and Veterinary Medicine, Usher Institute Weatherill, David; The University of Edinburgh College of Medicine and Veterinary Medicine Katikireddi, Srinivasa; University of Glasgow, MRC/CSO Social & Public Health Sciences Unit Kerr, Steven; The University of Edinburgh College of Medicine and Veterinary Medicine, Usher Institute Moore, Emily; Public Health Scotland, Data Driven Innovation Pesenti, Elisa; The University of Edinburgh, Usher Institute Quint, Jennifer; Imperial College London, Respiratory Epidemiology, Occupational Medicine and Public Health Shah, Syed Ahmar; The University of Edinburgh Usher Institute of Population Health Sciences and Informatics Shi, Ting; The University of Edinburgh College of Medicine and Veterinary Medicine, Usher Institute Simpson, Colin; Victoria University of Wellington Robertson, Chris; University of Strathclyde, Department of Mathematics and Statistics Sheikh, Aziz; The University of Edinburgh College of Medicine and Veterinary Medicine, Usher Institute
<b>Primary Subject Heading</b>:	Public health
Secondary Subject Heading:	Public health
Keywords:	PUBLIC HEALTH, COVID-19, Protocols & guidelines < HEALTH SERVICES ADMINISTRATION & MANAGEMENT

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



1  
2  
3  
4 **1 Deriving and validating a risk prediction model for long COVID-19:**  
5 **2 protocol for an observational cohort study using linked Scottish data**  
6  
7  
8

9 4 Luke Daines\*<sup>1</sup>, Rachel H Mulholland\*<sup>1</sup>, Eleftheria Vasileiou<sup>1</sup>, Vicky Hammersley<sup>1</sup>, David  
10 5 Weatherill<sup>1</sup>, Srinivasa Vittal Katikireddi<sup>2</sup>, Steven Kerr<sup>1</sup>, Emily Moore<sup>3</sup>, Elisa Pesenti<sup>4</sup>, Jennifer  
11 6 Quint<sup>5</sup>, Syed Ahmar Shah<sup>1</sup>, Ting Shi<sup>1</sup>, Colin R Simpson<sup>1,6</sup>, Chris Robertson<sup>3,7</sup>, Aziz Sheikh<sup>1</sup>  
12  
13  
14  
15

16 8 \*Contributed equally  
17  
18  
19

- 20  
21 10 1. Usher Institute, University of Edinburgh, Edinburgh, UK  
22 11 2. MRC/CSO Social & Public Health Sciences Unit, University of Glasgow, Glasgow, UK  
23 12 3. Public Health Scotland, Glasgow and Edinburgh, UK  
24 13 4. Institute of Cell Biology, University of Edinburgh, Edinburgh, UK  
25 14 5. Faculty of Medicine, National Heart & Lung Institute, Imperial College London, London,  
26 15 UK  
27 16 6. School of Health, Wellington Faculty of Health, Victoria University of Wellington,  
28 17 Wellington, NZ  
29 18 7. Department of Mathematics and Statistics, University of Strathclyde, Glasgow, UK  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39

40 20 **Corresponding author**

41 21 Dr Luke Daines

42 22 Usher Institute, University of Edinburgh,

43 23 Doorway 3, Old Medical School, Teviot Place, Edinburgh United Kingdom

44 24 Email: luke.daines@ed.ac.uk  
45  
46  
47  
48  
49  
50  
51

52 26 **Word count:** 3944 / 4000  
53  
54  
55  
56  
57  
58  
59  
60

# 1 **ABSTRACT**

## 2 **Introduction**

3 Coronavirus disease 2019 (COVID-19) is commonly experienced as an acute illness, yet  
4 some people continue to have symptoms that persist for weeks, or months (commonly  
5 referred to as “long-COVID”). It remains unclear which patients are at highest risk of  
6 developing long-COVID. In this protocol, we describe plans to develop a prediction model to  
7 identify individuals at risk of developing long-COVID.

## 8 **Methods and analysis**

9 We will use the national Early Pandemic Evaluation and Enhanced Surveillance of COVID-  
10 19 (EAVE II) platform, a population-level linked dataset of routine electronic healthcare data  
11 from 5.4 million individuals in Scotland. We will identify potential indicators for long-COVID  
12 by identifying patterns in primary care data linked to information from out-of-hours GP  
13 encounters, accident and emergency visits, hospital admissions, outpatient visits, medication  
14 prescribing/dispensing, and mortality. We will investigate the potential indicators of long-  
15 COVID by performing a matched analysis between those with a positive reverse  
16 transcriptase polymerase chain reaction (RT-PCR) test for Severe Acute Respiratory  
17 Syndrome 2 coronavirus (SARS-CoV-2) infection and two control groups; 1) individuals with  
18 at least one negative RT-PCR test and never tested positive; 2) the general population  
19 (everyone who did not test positive) of Scotland. Cluster analysis will then be used to  
20 determine the final definition of the outcome measure for long-COVID. We will then derive,  
21 internally and externally validate a prediction model to identify the epidemiological risk  
22 factors associated with long-COVID.

## 23 **Ethics and dissemination**

24 The EAVE II study has obtained approvals from the Research Ethics Committee (reference:  
25 12/SS/0201), and the Public Benefit and Privacy Panel for Health and Social Care  
26 (reference: 1920-0279). Study findings will be published in peer-reviewed journals and  
27 presented at conferences. Understanding the predictors for long-COVID and identifying the  
28 patient groups at greatest risk of persisting symptoms will inform future treatments and  
29 preventative strategies for long-COVID.

30  
31 **Word count:** 296/300

## Strengths and limitations of this study

- We will use national data on ~99% of the Scottish population using the Early Pandemic Evaluation and Enhanced Surveillance of COVID-19 (EAVE II) platform.
- Our study will be unable to identify long-COVID patients who have not been in contact with healthcare services in Scotland.
- Identifying long-COVID using routinely collected electronic health records may be challenging due to the lack of a standardised definition and variation in coding practices across healthcare systems.
- To improve the identification of long-COVID (and associated clinical features) we intend to use free text in addition to the coded data available in electronic health records.
- We are actively involving individuals who have experienced long-COVID to shape the research and ensure relevance to patients and the public.



## 1 INTRODUCTION

2 In December 2019, an outbreak of a novel coronavirus was reported in Wuhan, China. The  
3 World Health Organization (WHO) declared the outbreak a global pandemic named  
4 coronavirus disease 2019 (COVID-19) caused by the Severe Acute Respiratory Syndrome 2  
5 (SARS-CoV-2) coronavirus. By November 2021, the WHO had reported over 240 million  
6 confirmed cases and at least five million deaths worldwide,[1] with more than nine million  
7 confirmed cases and over 140,000 deaths reported in the United Kingdom (UK).[2]

8 The severity and duration of the acute SARS-CoV-2 infection varies widely. Most people are  
9 asymptomatic or experience mild-to-moderate symptoms, while a smaller proportion (10-  
10 15%) of cases experience more severe illness.[2] The majority of people recover after two to  
11 six weeks depending on disease severity.[3] However, some individuals have symptoms that  
12 last or recur for weeks or months after the initial acute infection.[3-18] Long-term effects of  
13 COVID-19 can present with a wide range of clinical features, relating to cardiovascular,  
14 neurological, respiratory and other organ systems, including mental health.[3-18] Common  
15 symptoms include fatigue, breathlessness, headaches, muscle weakness, joint pain and loss  
16 of taste or smell.[3,7,8,15-18]

17 Unified guidance to manage the long-term effects of COVID-19 in the UK has been  
18 developed by the National Institute for Health and Care Excellence (NICE), Scottish  
19 Intercollegiate Guidelines Network (SIGN) and the Royal College of General Practitioners  
20 (RCGP).[16] The guidance described two working case definitions of ongoing symptomatic  
21 COVID-19 (individuals with signs and symptoms of COVID-19 from four weeks to 12 weeks)  
22 and post-COVID-19 syndrome (individuals with signs and symptoms that develop during or  
23 following an infection consistent with COVID-19, continue for more than 12 weeks and are  
24 not explained by an alternative diagnosis).[16] The term 'long-COVID' therefore commonly  
25 refers to those who continue to present signs and symptoms four weeks after acute COVID-  
26 19 infection i.e. both ongoing symptomatic COVID-19 and post-COVID-19 syndrome.[16]

27 In Scotland, patients with symptoms suggestive of long-COVID are advised to seek medical  
28 care from their general practitioner (GP).[19] Diagnostic codes (Read codes, version 2)  
29 within the Scottish GP electronic system were introduced in March 2021 using NICE-led  
30 working definitions of long-COVID.[20] Equivalent diagnostic codes were also introduced in  
31 Scotland's Scottish Clinical Coding Standards using International Classification of Diseases  
32 10<sup>th</sup> Revision (ICD-10) codes within secondary care data in February 2021.[21] The long-  
33 COVID diagnostic codes are available in the supplementary material.

34 Despite the progress in diagnostic coding, the prevalence and risk factors associated with  
35 long-COVID remain poorly understood, reflecting the lack of an agreed operational definition,

1 the absence of diagnostic tests and the considerable variation in presentation. Reviews on  
2 the long-COVID literature have found that it was difficult to estimate the prevalence of  
3 persistent COVID-19 symptoms with certainty.[17,18] Therefore, alternative methods need to  
4 be adopted to identify those with long-COVID, so that the long-term consequences of  
5 COVID-19 illness can be better understood and individuals at highest risk of developing  
6 long-COVID can be identified early.[3,22-24] In this study, we aim to derive and validate a  
7 risk prediction model to estimate the probability that an individual will develop long-COVID.  
8 Our objectives are to: i) create an operational definition of long-COVID through studying  
9 health system interactions using a national linked healthcare dataset; ii) derive and validate  
10 a risk prediction model to estimate the probability of developing long-COVID; and iii)  
11 enhance the risk prediction model using machine learning.

## 12 **METHODS AND ANALYSIS**

### 13 **Study design and population**

14 We will undertake a national prospective population-based cohort study using the national  
15 Early Pandemic Evaluation and Enhanced Surveillance of COVID-19 (EAVE II)  
16 platform.[25,26] EAVE II comprises of routinely collected primary care, secondary care,  
17 laboratory and serology data from 5.4 million Scottish residents registered with a GP (~99%  
18 of the Scottish population) from February 2020.[25,26] We will primarily focus on adults  
19 (aged  $\geq 18$  years) but will consider extending the cohort to include children (aged  $<18$  years)  
20 if there are sufficient numbers of individuals in this age group. We intend to utilise data from  
21 February 2020 up to March 2023. The study started on 1 March 2021 and is scheduled to  
22 end on 28 February 2023.

### 23 **Inclusion/exclusion criteria**

24 Since the baseline population for this study is everyone registered with a GP, those who are  
25 not registered with a GP in Scotland will be excluded from the analyses.

### 26 **Sample size calculation**

27 We are using the whole population of Scotland and therefore sample size calculations are not  
28 applicable.

### 29 **Databases**

30 The EAVE II platform links a wide range of routine healthcare datasets using  
31 pseudonymised identifiers of National Health Service (NHS) Scotland's Community  
32 Healthcare Index (CHI). We will use these routinely collected data sources (described below)  
33 to identify individuals with long-COVID and to determine their characteristics in the EAVE II  
34 cohort.

## 1 Primary care data

2 Primary care data will be extracted from GP practices via EAVE II's trusted third party  
3 Albasoft Ltd.[25,26] GPs in the UK provide healthcare services that are free at the point of  
4 service and usually act as the first point of contact into the healthcare system. This data  
5 source captures all clinical and administrative activity at GPs and the characteristics of  
6 registered patients. These data are stored either as: 1) clinical codes; or 2) written free  
7 text.[27] The latter is used to capture detailed information on any encounter and may provide  
8 additional information not available in coded data. In order to include data from primary care  
9 encounters when GP practices are closed, we will use out-of-hours (OOH) records derived  
10 from the Public Health Scotland (PHS) Primary Care OOH Data Mart.[26]

## 11 Secondary care data

12 Activity in hospital-based care will be extracted from the Scottish Morbidity Record (SMR) 01  
13 which holds detailed information on hospital admissions, such as the specific area of clinical  
14 activity (specialty), the facility of care, patient management and new diagnoses.[28]  
15 Diagnoses in SMR01 will be extracted using ICD-10 codes.[28] For data on intensive care,  
16 we will use the Scottish Intensive Care Society Audit Group (SICSAG) dataset of all adult  
17 patients admitted to Intensive Care Units (ICU) and High Dependency Units (HDU) in  
18 Scotland.[28] For outpatient care, we will use the SMR00 dataset, which captures outpatient  
19 activity in specialist clinics such as physiotherapy.[28]

## 20 Laboratory data

21 All COVID-19 testing will be obtained from the Electronic Communication of Surveillance in  
22 Scotland (ECOSS) dataset. This surveillance data contains all reverse transcriptase  
23 polymerase chain reaction PCR (RT-PCR) tests, carried out in Scotland.[26] Sequencing  
24 data will be obtained from the Centre of Genomics (COG) and will make it possible to  
25 account for the variant of SARS-CoV-2 during model building.

## 26 Vaccination data

27 COVID-19 vaccination data, including vaccination type and number of doses administered,  
28 will be available from two sources: GP records and the Turas Vaccination Management Tool  
29 (TVMT), a web-based tool used to record community vaccinations in Scotland.[29]

1  
2  
3 1 Telehealth data  
4

5  
6 2 Telehealth in Scotland is operated by NHS 24 Scotland, which delivers telephone and online  
7 3 services [30]. We are specifically interested in the NHS 24 111 teleservice, which provides  
8 4 OOH advice. During the pandemic, this service was expanded to include a COVID-19  
9 5 helpline which was used to provide advice and triage patients to COVID-19 Assessment  
10 6 Centres.[30]  
11  
12  
13

14  
15 7 Prescribing data  
16

17 8 Prescription data relating to all medications prescribed and dispensed in the community in  
18 9 Scotland will be extracted from the Prescribing Information System (PIS).[26] These  
19 10 medications are coded using the British National Formulary (BNF) code lists.[31] For  
20 11 medication data within hospitals, Hospital Electronic Prescribing and Medicines  
21 12 Administration (HEPMA) which are available for five Health Boards will be used.[26]  
22  
23  
24  
25

26 13 Mortality data  
27

28  
29 14 Mortality data will be taken from death registry data within the National Records of Scotland.  
30 15 These records hold information included on the death certificate, including cause(s) of death  
31 16 which are recoded using ICD-10 codes.[26]  
32  
33  
34

35 17 Other data  
36

37 18 We will explore the use of other linkages available within the EAVE II platform. These  
38 19 include Scotland's Census 2011 from NHS Research Scotland (NRS) for information on  
39 20 ethnicity, disability, and occupation as part of the EAVE II sub-study for ethnic and social  
40 21 inequalities in COVID-19 outcomes in Scotland.[32] We will also consider linkages and  
41 22 comparisons to Generation Scotland's CovidLife surveys which launched in April 2020 to  
42 23 capture how COVID-19 has been affecting volunteers in the UK.[33]  
43  
44  
45  
46  
47

48 24 **Determining an operational definition for long-COVID**

49 25 We will base our operational definition on the case definitions for the effects of COVID-19  
50 26 illness at different time periods developed by NICE[16]:  
51

- 52  
53 27 1. Acute COVID-19 infection: individuals with signs and symptoms of COVID-19 for up  
54 28 to 4 weeks  
55  
56 29 2. Ongoing symptomatic COVID-19: individuals with signs and symptoms of COVID-19  
57 30 from 4-12 weeks  
58  
59  
60

- 1  
2  
3 1 3. Post-COVID-19 syndrome: individuals with signs and symptoms that develop during  
4 2 or following an infection consistent with COVID-19, continue for more than 12 weeks  
5 3 and are not explained by an alternative diagnosis. The post-COVID-19 syndrome  
6 4 usually presents with clusters of symptoms, often overlapping, which can fluctuate  
7 5 and change over time and can affect any organ system.

8 6 Long-COVID commonly refers to those who continue to present with signs and symptoms  
9 7 four or more weeks after acute COVID-19 infection, therefore our primary outcome will  
10 8 include both ongoing symptomatic COVID-19 and post-COVID-19 syndrome. Our secondary  
11 9 outcome will focus on the clinical encounters suggestive of the post-COVID-19 syndrome.  
12 10 Further details are in the statistical analyses.

### 11 **Population characteristics**

12 12 Population characteristics will be explored to assess the risk factors for developing long-  
13 13 COVID and to account for any confounding in our analyses.

#### 14 Socio-demographics

15 15 Age will be determined based on the available GP data and will be available as a continuous  
16 16 and categorical variable. Those aged over 100 will be truncated into the one group to  
17 17 overcome low sample size issues. Sex at birth will be included as a binary variable  
18 18 (female/male). Deprivation status will be derived from the Scottish Index of Multiple  
19 19 Deprivation (SIMD) 2020 quintile of the resident's postcode associated with their GP  
20 20 registration. Ethnicity data will also be included if completeness and quality of data is  
21 21 adequate. We will also consider other available information such as Body Mass Index (BMI)  
22 22 and smoking status (smoker, ex-smoker, non-smoker and unknown).

#### 23 Geographical

24 24 Area of residence in terms of NHS Scotland Health Boards and local authorities will be  
25 25 considered. Settlement type will be determined by the urban/rural 6-fold classification (UR6).  
26 26 Type of residence will also be considered such as private residence, care home and  
27 27 social/council housing if data are available.

#### 28 Clinical characteristics

29 29 Using diagnostic codes from the QCOVID algorithm,[34] we will identify the following  
30 30 conditions: a) cardiovascular; b) diabetes (type 1 and type 2); c) respiratory d) cancer (blood  
31 31 cancer, chemotherapy, lung or oral cancer, marrow transplant, radiotherapy); e)

1  
2  
3 1 neurological; f) other conditions, such as liver cirrhosis, osteoporotic fracture, rheumatoid  
4 2 arthritis, systemic lupus erythematosus, sickle cell disease, venous thromboembolism, solid  
5 3 organ transplant, renal failure (chronic kidney disease stages 3-5 with or without dialysis or  
6 4 transplant).[34]  
7  
8  
9

10 5 Severity of acute COVID-19 illness  
11  
12

13 6 Admission to hospital, any requirement for treatment in the ICU, and death will be used to  
14 7 categorise the severity of COVID-19 infection. We will define a COVID-19 hospitalisation as  
15 8 a RT-PCR confirmed positive test for SARS-CoV-2 in the 28 days prior to admission, or  
16 9 admission with an ICD-10 code for COVID-19. A COVID-19 ICU admission will be defined as  
17 10 a RT-PCR confirmed positive test for SARS-CoV-2 in the 28 days prior to ICU admission. A  
18 11 COVID-19 death will be defined as dying within 28 days of confirmed or probable COVID-19.  
19  
20  
21  
22

## 23 12 **Missing data**

24  
25 13 The amount of missing data will be examined for each variable of interest. Continuous  
26 14 variables, for example BMI, will be imputed using predictive mean matching or imputation by  
27 15 chained equations if appropriate. Categorical variables with missing data will have a distinct  
28 16 group of 'Unknown'. We will consider dealing with these missing categorical variables by  
29 17 either keeping the distinct group, imputing them using chained equations or removing them  
30 18 from the analysis. The latter will be a complete case analysis, which will reduce the total  
31 19 sample size.  
32  
33  
34  
35

## 36 20 **Statistical analyses**

37  
38  
39 21 Developing an operational definition of long-COVID  
40  
41

42 22 We will firstly derive an operational definition for long-COVID by identifying patterns in  
43 23 clinical interactions within NHS Scotland services that may suggest long-COVID. A visual  
44 24 illustration of our intended methods is shown in Figure 1.  
45  
46  
47

### 48 25 *Indicators of long-COVID*

49  
50 26 We are interested in a) GP interactions; b) hospital admissions; c) outpatient attendances; d)  
51 27 A&E visits; e) OOH encounters; f) NHS 24 telehealth interactions; g) medications (from GP  
52 28 prescribing and primary care pharmacy dispensing data); and h) all-cause mortality. Our  
53 29 primary focus will be on the GP data, with other healthcare data providing corroborative  
54 30 information (Step 1, Figure 1). Information within these electronic health datasets will serve  
55 31 as an investigative list of potential indicators for long-COVID.  
56  
57  
58  
59  
60



1  
2  
3 1 For the healthcare services (sources a to f), we will investigate the frequency of interactions  
4 2 and the reasons for each interaction which will include any new diagnoses (categorised by  
5 3 body system), treatments, tests or procedures related to long-COVID. For medications (g),  
6 4 we will investigate the frequency and type of new prescriptions using British National  
7 5 Formulary (BNF) chapters. For all-cause mortality (h) we will record the causes of death  
8 6 using ICD-10 codes. Figure 2 summarises the different data sources and potential indicators  
9 7 of long-COVID we intend to investigate. The dataset will comprise of categorical binary  
10 8 variables (e.g., diagnosis or not) and numerical variables (e.g., number of consultations).

11 9 GP records provide a rich source of primary care data, with the coded data providing  
12 10 additional context to interactions such as the type of interaction (e.g., consultation,  
13 11 encounter, remote or face-to-face), referrals to specialty care and sick notes. For more  
14 12 detailed information on signs and symptoms that are indicative of long-COVID, we intend to  
15 13 use written free text available from GP records. We will use natural language processing  
16 14 (NLP) to identify key words or phrases of signs and symptoms relating to long-COVID. This  
17 15 NLP model will be applied to all written free text using Computer-Assisted (diagnostic)  
18 16 Coding (CAC). This will create derived codes associated with the key words and phrases (1  
19 17 if the text mentions word or phrase, 0 otherwise).[35,36] These derived codes will be treated  
20 18 in a similar way to GP codes as discussed above. Figure 3 demonstrates this process of  
21 19 transforming the written GP free text into derived codes.

22 20 To initially explore potential long-COVID indicators, we will obtain summary level counts of  
23 21 all codes of interest within these healthcare datasets on individuals who tested positive and  
24 22 negative for COVID-19 using a RT-PCR test. We will count the frequency of these data  $\geq 4$   
25 23 weeks after the date of the test. This will inform which codes relating to potential long-COVID  
26 24 indicators will be extracted on a patient level for further analysis.

### 25 *Matched analysis.*

26 26 To identify which of these long-COVID indicators are most important, we will perform a  
27 27 matched analysis (Step 2, Figure 1). The exposed group will be defined as the first date an  
28 28 individual tested positive for COVID-19 using a RT-PCR test. Two control groups will be  
29 29 assigned: 1) individuals who have had at least one negative RT-PCR test and have never  
30 30 tested positive up to the date of the exposed match testing positive; and 2) the general  
31 31 population (everyone who did not test positive) of Scotland. These control groups will be  
32 32 investigated in turn.

33 33 We will use risk-set matching in a 3:1 ratio by time-varying propensity score matching. This  
34 34 will be based on the likelihood of testing positive for COVID-19 and will consider

1  
2  
3 1 incorporating the following characteristics: sex, age, geography, comorbidities, risk factors,  
4 2 number of previous SARS-CoV-2 tests, deprivation status and urban-rural settlement. The  
5 3 adequacy of the matching will be assessed by checking for imbalance of the individual  
6 4 covariates across exposure groups.

7 5 For the matched analysis, each potential long-COVID indicator will be treated as its own  
8 6 dependent variable in turn. Follow-up will begin from four weeks after the exposed tested  
9 7 positive for COVID-19. Follow-up will end on either the date of event (if indicator is a binary  
10 8 variable), the control testing positive for COVID-19, death from any cause or the end of the  
11 9 follow-up period. Controls who have a positive test will be eligible to be included in the  
12 10 exposed group.

13 11 The long-COVID indicators will be compared between the exposed and control groups using  
14 12 statistical tests such as two-sample proportions test (for binary indicators), two-sample t-  
15 13 tests (for continuous indicators), Kaplan-Meier curves to inspect cumulative incidence, and  
16 14 survival analysis to look at the potential impact of interventions on long-COVID symptoms.

17 15 We also plan to conduct similar analyses on the whole cohort without propensity score  
18 16 matching. We will consider stratifying by age and sex if numbers allow.

### 19 17 *Cluster analysis*

20 18 Clusters of long-COVID presentations in the exposed group will be investigated further,  
21 19 using the long-COVID indicators as our clustering input (Step 3, Figure 1). The indicators will  
22 20 be summarised using a window of four or more weeks after initially testing positive (e.g., the  
23 21 number of interactions  $\geq 4$  weeks after the test). These indicators will not include the  
24 22 diagnostic codes for long-COVID since we are aiming to provide a more accurate alternative  
25 23 to this measurement of long-COVID.

26 24 We will explore both hierarchical clustering and k-means clustering, using distance  
27 25 measurements such as the Gower Distance which is a suitable measurement of similarity for  
28 26 mixed categorical and numeric data.[37] We will also investigate clusters based on latent  
29 27 class analysis. We will then internally validate these clusters using statistics such as the  
30 28 silhouette coefficient and the Dunn index.[38,39] Comparisons between the clusters and  
31 29 long-COVID diagnostic codes will be undertaken for validation. The final set of clusters of  
32 30 long-COVID indicators will serve as our operational definition for long-COVID.

### 33 31 *Sensitivity analyses*

34 32 We will perform a variety of sensitivity analyses to test the robustness of our long-COVID  
35 33 definition. This includes evaluating the start of follow-up to 12 weeks, to explore whether the



1  
2  
3 1 alternative outcome definition of 'post-COVID-19 syndrome' display different clinical  
4 2 pathways. We will also investigate the patterns in the long-COVID indicators associated with  
5 3 the diagnostic long-COVID codes. To capture those who may be suffering from long-COVID  
6 4 but did not formally test positive for COVID-19 (or tested positive on a lateral flow device  
7 5 only), we will investigate the long-COVID indicators in the general population. We will also  
8 6 stratify by time-period, for example during the different peaks of positive cases in Scotland  
9 7 (e.g., March 2020 to July 2020, August 2020 to April 2021, December 2021 to March  
10 8 2022).[2] This will also reflect the dominant COVID-19 variants during the different waves of  
11 9 infection.

#### 10 Deriving and validating a risk prediction model for long-COVID

11 We will use the transparent reporting of a multivariable prediction model for individual  
12 11 prognosis or diagnosis (TRIPOD) guidelines to report the derivation and validation of the  
13 12 long-COVID prediction model (see completed checklist in the supplementary material).[40]  
14 13 The model will be derived using data from everyone in the cohort (defined above) who  
15 14 received a positive PCR test. We acknowledge the cohort may not include all people who  
16 15 had COVID-19 (for instance those who only tested positive by lateral flow device) but a  
17 16 positive PCR result is the most reliable marker of COVID-19 available from national  
18 17 datasets.

#### 19 *Descriptive analysis*

20 We will begin analysis by conducting descriptive analyses to visually inspect and summarise  
21 20 the types of potential long-COVID presentations within the clusters. Next, summaries of the  
22 21 geographical, sociodemographic and risk factor profile of those presenting with the long-  
23 22 COVID clusters will be reported.

#### 24 *Outcome*

25 The outcome for the risk prediction model will be the derived operational definition of long-  
26 25 COVID defined from the cluster analysis. This will be dependent on the number of optimum  
27 26 clusters and the different classifications of long-COVID presentation. Depending on the  
28 27 clusters, we will classify our outcome into a binary variable of belonging to one (or more)  
29 28 cluster(s) (1) or otherwise (0).

#### 30 *Predictor variables*

31 Predictors for the risk prediction model will consist of the patient characteristics, including  
32 31 information on socio-demographics, geographical, clinical comorbidities and severity of

1  
2  
3 1 COVID-19 infection. These will be a mixture of continuous, binary and categorical variables.  
4  
5 2 Continuous variables will be tested for linearity and for more flexible relationships (using  
6  
7 3 smooth splines). Groupings of continuous variables will be explored if necessary.  
8

9 4 *Type of model*

10  
11 5 We intend to use a multivariable logistic regression model.  
12  
13

14 6 *Selection of predictors*

15  
16  
17 7 We will build our model using stepwise selection based on the Akaike's Information Criterion  
18 (AIC) and Bayesian information criterion (BIC). To assess the fit of the model parameters,  
19 8 the maximum likelihood ratio test will also be used.  
20 9  
21

22  
23 10 *Model evaluation/performance*

24  
25 11 To evaluate the model's goodness of fit, we will use appropriate performance evaluation  
26 12 metrics such as the area under the ROC curve (captures the accuracy of the model  
27 13 discriminating between the outcome), and the calibration plot and slope (visualises the  
28 14 observed vs predicted values). Other evaluation measures such as the specificity (true  
29 15 negative rate), sensitivity (true positive rate) and accuracy will be considered if appropriate.  
30 16 We will also directly compare the predicted and observed values.  
31  
32  
33  
34  
35

36 17 *Model validation*

37  
38  
39 18 The model will be internally validated using k-fold cross validation. We will validate using  
40 19 different time periods as specified by one of the discussed sensitivity analyses in the  
41 20 clustering analysis. We will explore opportunities for external validation, comparison and  
42 21 meta-analysis with other long-COVID initiatives.  
43  
44  
45

46 22 *Risk groups*

47  
48  
49 23 To categorise the output of the model for further use by clinicians and COVID-19 patients,  
50 24 we will consider stratifying patients into risk groups based on the predictive probabilities in  
51 25 the multivariable model, for example three groups of low, moderate and high risk of long-  
52 26 COVID.  
53  
54  
55  
56  
57  
58  
59  
60

## 1 *Sensitivity analysis*

2 Sensitivity analyses for the risk prediction algorithm will depend on the outcomes from the  
3 sensitivity analyses from Objective 1 of developing an operational definition for long-COVID.

## 4 Enhancing the prediction model using machine learning

5 Advances in machine learning will be utilised to enhance the development and validation of  
6 the prediction model. Specifically, we will systematically explore the use of supervised  
7 learning algorithms such as penalised models (e.g., LASSO regression), naïve Bayes  
8 classifier, gradient boosting decision trees and random forests to further improve the  
9 prediction model developed with traditional statistical methods. We will also consider using  
10 ensemble learning methods to strategically combine multiple models to obtain better  
11 predictive performance.

## 12 **Patient and public involvement**

13 Lay input has shaped the development of this research and will continue throughout the  
14 project through the patient and public involvement (PPI) co-applicant, the EAVE II Public  
15 Advisory Group (PAG), and long-COVID Scotland. PPI members will collaborate with the  
16 research team to provide real world perspectives when analysing and interpreting study  
17 findings ensuring that the work considers the needs, interests and concerns of patient and  
18 public members.

## 19 **ETHICS AND DISSEMINATION**

### 20 Ethical approval

21 This study forms part of the EAVE II project which is investigating epidemiological risk  
22 factors of COVID-19 disease. All data will be anonymised before being made available to the  
23 research team. EAVE II has already obtained permissions from the Research Ethics  
24 Committee (REC reference: 12/SS/0201), NHS Research and Development, GPs, NHS  
25 Health Boards and the Public Benefit and Privacy Panel (PBPP) for Health and Social Care  
26 (reference number: 1920-0279).

### 27 Dissemination

28 To ensure the greatest impact of our findings, we will actively disseminate to three key  
29 audiences: policy/public health, academic and community-based.

1  
2  
3 1 *Policy and Public Health*  
4

5  
6 2 Findings from this project will provide evidence to help NHS Scotland and other international  
7 3 policy makers identify groups of the population who are at most risk of long-COVID and  
8 4 related complications. We will work with partners in NHS Scotland, Public Health Scotland  
9 5 (PHS) and the Scottish Government to establish the best ways of disseminating these  
10 6 results and influencing policy. We also plan to disseminate findings through the National  
11 7 Core Studies programme, a UK government initiative supported by Health Data Research  
12 8 UK (HDRUK), in partnership with the Office for National Statistics (ONS).  
13  
14  
15  
16  
17

18 9 *Academic*  
19

20 10 We will communicate our findings through presentations at major national and international  
21 11 scientific meetings and through publications in relevant peer-reviewed journals. We will  
22 12 publish our code and data dictionary on the EAVE II's GitHub repository  
23 13 (<https://github.com/EAVE-II>).  
24  
25  
26  
27

28 14 *Community-based*  
29

30 15 We will write lay summaries of publications and create infographics to further communicate  
31 16 our findings via press releases, public and patient engagement events, social media and the  
32 17 EAVE II website (<https://www.ed.ac.uk/usher/EAVE-ii>).  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## 1           2           3           4           5           6           7           8           9           10           11           12           13           14           15           16           17           18           19           20           21           22           23           24           25           26           27           28           29           30           31           32           33           34           35           36           37           38           39           40           41           42           43           44           45           46           47           48           49           50           51           52           53           54           55           56           57           58           59           60

### 1   **AUTHORS CONTRIBUTIONS**

2   AS conceived this manuscript. RM and LD led the writing of the manuscript with critical  
3   revision from EV, VH, DW, SVK, EM, EP, JKQ, TS and CRS. Statistical methods were  
4   reviewed by EAVE II's lead statistician CR and the other co-authors VK, LD, SAS and SK.  
5   All authors reviewed the manuscript and gave final approval to be published.

### 7   **FUNDING STATEMENT**

8   This work was supported by the Chief Scientist Office, grant number COV/LTE/20/15. EAVE  
9   II is supported by a grant (MC\_PC\_19075) from the Medical Research Council; a grant  
10  (MC\_PC\_19004) from BREATHE–The Health Data Research Hub for Respiratory Health,  
11  funded through the UK. Research and Innovation Industrial Strategy Challenge Fund and  
12  delivered through Health Data Research UK; a grant from the Data and Connectivity  
13  National Core Study, led by Health Data Research UK in partnership with the Office for  
14  National Statistics and funded by UK Research and Innovation (grant ref MC\_PC\_20058);  
15  Public Health Scotland; and the Scottish Government Director General for Health and Social  
16  Care. SVK acknowledges funding from a NRS Senior Clinical Fellowship (SCAF/15/02), the  
17  Medical Research Council (MC\_UU\_00022/2) and the Scottish Government Chief Scientist  
18  Office (SPHSU17).

### 19  **COMPETING INTERESTS STATEMENT**

20  AS is a member of the Scottish Government Chief Medical Officer's COVID-19 Advisory  
21  Group and its Standing Committee on Pandemics. He is a member of the UK Government's  
22  Risk Stratification Subgroup and Astra-Zeneca's Thrombotic Thrombocytopenic Taskforce.  
23  All roles are unremunerated. SVK was co-chair of the Scottish Government's Expert  
24  Reference Group on Ethnicity and COVID-19 and a member of the UK Government's  
25  Scientific Advisory Group on Emergencies (SAGE) subgroup on ethnicity. All other authors  
26  declare no competing interests.

### 27  **ACKNOWLEDGEMENTS**

28  We are grateful for the support from Public Health Scotland (PHS) and Albasoft Ltd on the  
29  development of the protocol and data extraction. Methods were also developed with support  
30  from the CONVALESCENCE study analysts. We would also like to acknowledge the PPI  
31  support from the EAVE II PAG members and long-COVID Scotland.

## 1 REFERENCES

- 1 2 1. World Health Organization. WHO Coronavirus Disease (COVID-19)  
3 Dashboard. Available from: <https://covid19.who.int/> (Accessed November  
4 2021)
- 5 2. Public Health Scotland. COVID-19 in Scotland Daily Dashboard. Available  
6 from: [https://public.tableau.com/profile/phs.covid.19#!/vizhome/COVID-  
7 19DailyDashboard\\_15960160643010/Overview](https://public.tableau.com/profile/phs.covid.19#!/vizhome/COVID-19DailyDashboard_15960160643010/Overview) (Accessed November  
8 2021)
- 9 3. World Health Organization. What we know about Long-term effects of  
10 COVID-19. Available from: [https://www.who.int/docs/default-  
11 source/coronaviruse/risk-comms-  
12 updates/update54\\_clinical\\_long\\_term\\_effects.pdf?sfvrsn=3e63eee5\\_8](https://www.who.int/docs/default-source/coronaviruse/risk-comms-updates/update54_clinical_long_term_effects.pdf?sfvrsn=3e63eee5_8)  
13 (Accessed November 2021)
- 14 4. Rayner C, Lokugame AU, Molokhia M. Covid-19: Prolonged and relapsing  
15 course of illness has implications for returning workers. *BMJ* 2020.  
16 Available from: [https://blogs.bmj.com/bmj/2020/06/23/covid-19-prolonged-  
17 and-relapsing-course-of-illness-has-implications-for-returning-workers/  
18 \(Accessed November 2021\)](https://blogs.bmj.com/bmj/2020/06/23/covid-19-prolonged-and-relapsing-course-of-illness-has-implications-for-returning-workers/)
- 19 5. Carfi A, Bernabei R, Landi F. Persistent Symptoms in Patients After Acute  
20 COVID-19. *JAMA* 2020;324(6):603-605.
- 21 6. del Rio C, Collins LF, Malani P. Long-term Health Consequences of  
22 COVID-19. *JAMA* 2020;324(17):1723-4.
- 23 7. Huang C, Huang L, Wang Y, et al. 6-month consequences of COVID-19 in  
24 patients discharged from hospital: a cohort study. *Lancet*  
25 2020;397(10270):220-32
- 26 8. Evans RA, McAuley H, Harrison EM, et al. Physical, cognitive, and mental  
27 health impacts of COVID-19 after hospitalisation (PHOSP-COVID): a UK  
28 multicentre, prospective cohort study. *Lancet Respir Med*. 2021;  
29 9(11):1275-1287 DOI: [https://doi.org/10.1016/S2213-2600\(21\)00383-0](https://doi.org/10.1016/S2213-2600(21)00383-0)
- 30 9. Ayoubkhani D, Khunti K, Nafilyan V, et al. Epidemiology of post-COVID  
31 syndrome following hospitalisation with coronavirus: a retrospective cohort  
32 study. medRxiv 2021.01.15.21249885; doi:  
33 <https://doi.org/10.1101/2021.01.15.21249885>
- 34 10. Mitrani RD, Dabas N, Goldberger JJ. COVID-19 cardiac injury:  
35 Implications for long-term surveillance and outcomes in survivors. *Heart*  
36 *Rhythm* 2020;17(11):1984-1990.
- 37 11. Mao L, Jin H, Wang M et al. Neurologic Manifestations of Hospitalized  
38 Patients With Coronavirus Disease 2019 in Wuhan, China. *JAMA*  
39 *Neurol* 2020;77(6):683-690.
- 40 12. George PM, Wells AU, Jenkins RG. Pulmonary fibrosis and COVID-19:  
41 the potential role for antifibrotic therapy. *Lancet Respir Med*  
42 2020;8(8):807-815.
- 43 13. Rajkumar RP. COVID-19 and mental health: A review of the existing  
44 literature. *Asian J Psychiatr* 2020;52:102066.
- 45 14. ZOE app: One in 20 people likely to suffer from 'Long COVID', but who  
46 are they? Available from: <https://covid.joinzoe.com/post/long-covid>  
47 (Accessed November 2021)



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

15. Office for National Statistics, Prevalence of ongoing symptoms following coronavirus (COVID-19) infection in the UK: 4 November 2021. Available from:  
<https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/prevalenceofongoingsymptomsfollowingcoronaviruscovid19infectionintheuk/4november2021> (Accessed November 2021)
16. NICE guideline. COVID-19 rapid guideline: managing the long-term effects of COVID-19. Available from:  
<https://www.nice.org.uk/guidance/ng188> (Accessed November 2021)
17. NIHR: Living with Covid19 – Second review. Available from:  
<https://evidence.nihr.ac.uk/themedreview/living-with-covid19-second-review/> (Accessed November 2021)
18. Usher Network for COVID-19 Evidence Reviews. What is the prevalence of the post-COVID-19 syndrome? Available from:  
[https://www.ed.ac.uk/files/atoms/files/uncover\\_028-01\\_summary\\_prevalence\\_of\\_post\\_covid-19\\_syndrome\\_2april.pdf](https://www.ed.ac.uk/files/atoms/files/uncover_028-01_summary_prevalence_of_post_covid-19_syndrome_2april.pdf) (Accessed November 2021)
19. NHS Inform. Longer-term effects of COVID-19 (long COVID).  
<https://www.nhsinform.scot/illnesses-and-conditions/infections-and-poisoning/coronavirus-covid-19/coronavirus-covid-19-longer-term-effects-long-covid> (Accessed November 2021)
20. Scottish Government. Management and recording of the long-term effects of COVID-19 (Long COVID). March 2020. Available from:  
<https://www.scimp.scot.nhs.uk/wp-content/uploads/CMO-Letter-09.03.21.pdf> (Accessed November 2021)
21. Public Health Scotland. Scottish Clinical Coding Standards. Number 27.  
<https://www.isdscotland.org/products-and-services/terminology-services/clinical-coding-guidelines/Docs/Scottish-clinical-coding-standards-Feb-2021-No-27.pdf> (Accessed November 2021)
22. Yelin D, Wirtheim E, Vetter P, et al. Long-term consequences of COVID-19: research needs. *Lancet Infect Dis* 2020;20(10):1115-1117.
23. Marshall M. The lasting misery of coronavirus long-haulers. *Nature*. 2020;585(7825):339-341.
24. Rimmer A. Covid-19: Impact of long term symptoms will be profound, warns BMA. *BMJ* 2020;370:m3218.
25. Simpson CR, Robertson C, Vasileiou E, et al. Early Pandemic Evaluation and Enhanced Surveillance of COVID-19 (EAVE II): protocol for an observational study using linked Scottish national data. *BMJ Open* 2020;10:e039097.
26. 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. DOI: <https://doi.org/10.1093/ije/dyab028>
27. Scottish Clinical Information Management in Practice (SCIMP). SCIMP Guide to Read Codes. Available from:  
<https://www.scimp.scot.nhs.uk/better-information/clinical-coding/scimp-guide-to-read-codes> (Accessed November 2021)

- 1  
2  
3 1 28. Information Services Division. SMR Datasets. General Clinical  
4 2 Information. Available from: [https://www.ndc.scot.nhs.uk/Data-](https://www.ndc.scot.nhs.uk/Data-Dictionary/SMR-Datasets/General-Clinical-Information/)  
5 3 [Dictionary/SMR-Datasets/General-Clinical-Information/](https://www.ndc.scot.nhs.uk/Data-Dictionary/SMR-Datasets/General-Clinical-Information/) (Accessed  
6 4 November 2021)
- 8 5 29. Turas. Turas Vaccination Management tool. Available from:  
9 6 <https://learn.nes.nhs.scot/42708/turas-vaccination-management-tool>  
10 7 (Accessed November 2021)
- 11 8 30. NHS24. Available from: <https://www.nhs24.scot/111/when-to-phone-111/>  
12 9 (Accessed November 2021)
- 14 10 31. National Institute for Health and Care Excellence (NICE). BNF. June  
15 11 2021. Available from: <https://bnf.nice.org.uk/> (Accessed November 2021)
- 17 12 32. University of Glasgow. Exploring Inequalities in COVID-19 in Scotland.  
18 13 Available from:  
19 14 [https://www.gla.ac.uk/researchinstitutes/healthwellbeing/research/mrccso-](https://www.gla.ac.uk/researchinstitutes/healthwellbeing/research/mrccso-socialandpublichealthsciencesunit/aboutus/covid19/inequalitiescovid/)  
20 15 [socialandpublichealthsciencesunit/aboutus/covid19/inequalitiescovid/](https://www.gla.ac.uk/researchinstitutes/healthwellbeing/research/mrccso-socialandpublichealthsciencesunit/aboutus/covid19/inequalitiescovid/)  
21 16 (Accessed November 2021)
- 23 17 33. Generation Scotland. CovidLife. Available from:  
24 18 <https://www.ed.ac.uk/generation-scotland/for-researchers/covidlife>  
25 19 (Accessed November 2021)
- 26 20 34. Cliff AK, Coupland CA, Keogh RH, et al. Living risk prediction algorithm  
27 21 (QCOVID) for risk of hospital admission and mortality from coronavirus 19  
28 22 in adults: national derivation and validation cohort study. *BMJ*  
29 23 2020;371:m3731.
- 31 24 35. Nguyen AN, Truran D, Kemp M, et al. Computer-Assisted Diagnostic  
32 25 Coding: Effectiveness of an NLP-based approach using SNOMED CT to  
33 26 ICD-10 mappings. *AMIA Annu Symp Proc.* 2018:807-816.
- 35 27 36. Koleck TA, Dreisbach C, Bourne PE, et al. Natural language processing of  
36 28 symptoms documented in free-text narratives of electronic health records:  
37 29 a systematic review. *J Am Med Inform Assoc.* 2019;26(4):364-379
- 38 30 37. Gower JC. A General Coefficient of Similarity and Some of Its Properties.  
39 31 *Biometrics* 1971;1:857-871.
- 41 32 38. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and  
42 33 validation of cluster analysis. *J Comput Appl Math* 1987;20:53-65.
- 43 34 39. Dunn J.C. A Fuzzy Relative of the ISODATA Process and Its Use in  
44 35 Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*  
45 36 1973;3(3)32-57, DOI: 10.1080/01969727308546046
- 47 37 40. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a  
48 38 multivariable prediction model for Individual Prognosis or Diagnosis  
49 39 (TRIPOD): explanation and elaboration. *Annals of internal medicine.*  
50 40 2015;162(1):W1-73.



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1 **FIGURE LEGENDS**

2 Figure 1: Schematic diagram of methods for developing an operational definition for Long  
3 COVID.

4 Figure 2: Data linkage diagram of long-COVID indicators and their data sources within the  
5 EAVE II cohort.

6 Figure 3: Schematic diagram of Natural Language Processing (NLP) and Computer Assisted  
7 Coding (CAC) on GP free text

8  
9  
10  
11  
12  
13

For peer review only

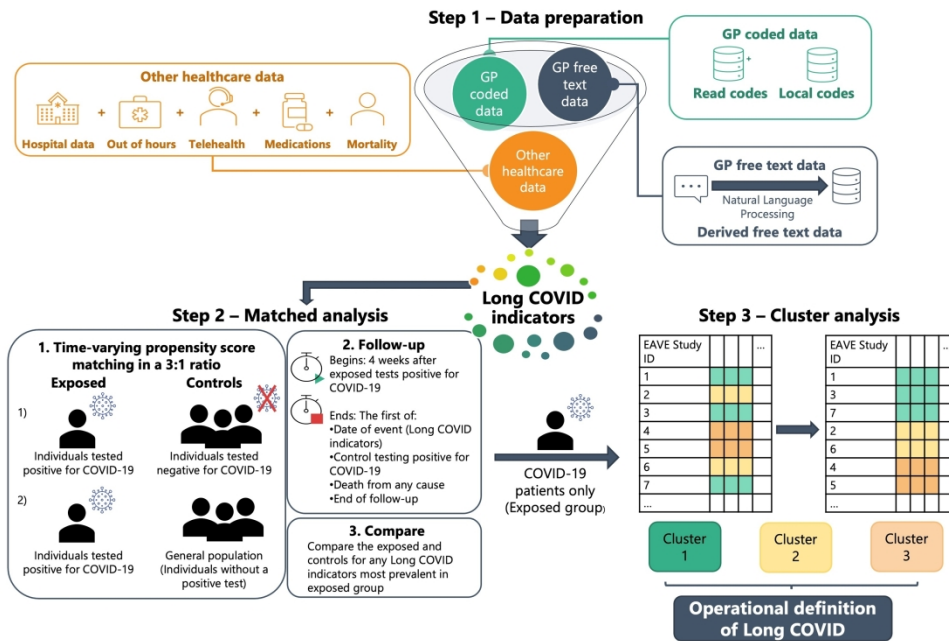


Figure 1: Schematic diagram of methods for developing an operational definition for Long COVID.

275x190mm (300 x 300 DPI)

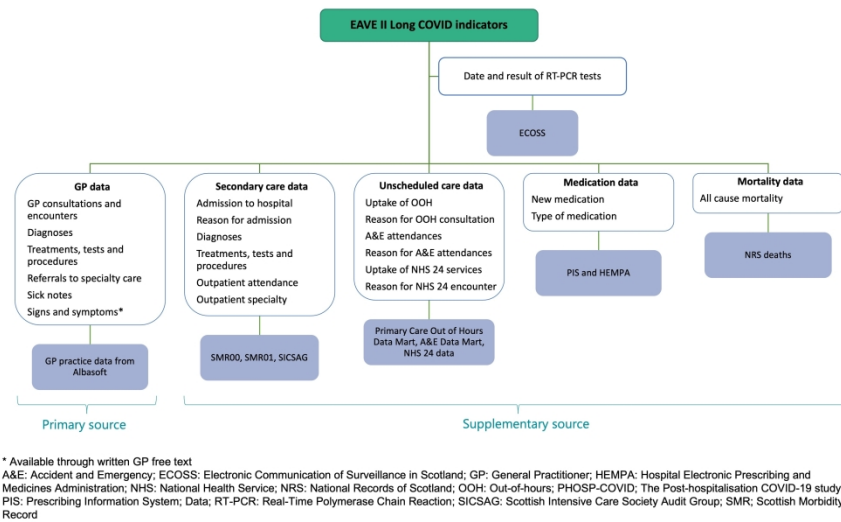
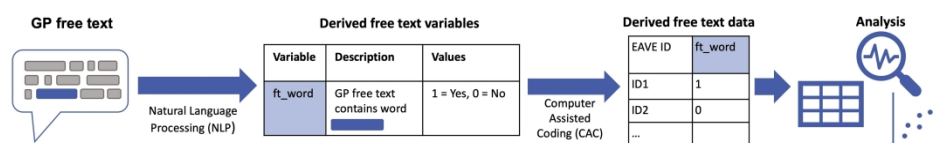


Figure 2: Data linkage diagram of long-COVID indicators and their data sources within the EAVE II cohort.

384x212mm (300 x 300 DPI)



13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Figure 3: Schematic diagram of Natural Language Processing (NLP) and Computer Assisted Coding (CAC) on GP free text

337x62mm (300 x 300 DPI)

# Deriving and validating a risk prediction model for long COVID-19: protocol for an observational cohort study using linked Scottish data

## SUPPLEMENTARY FILE

Diagnostic codes for Long COVID

*Local codes introduced in Scotland<sup>1</sup>*

Clinical Computer System	Code	Description
EMIS PCS	^ESCT1348648	Ongoing symptomatic COVID-19
EMIS PCS	^ESCT1348645	Post-COVID-19 syndrome
Vision	A7955	Ongoing symptomatic COVID-19
Vision	AyuJC	Post-COVID-19 syndrome

*ICD-10 emergency use codes for conditions related to COVID-19<sup>2</sup>*

ICD-10 Code	Description
U07.3	Personal history of COVID-19
U07.4	Post-COVID-19 condition
U07.5	Multisystem inflammatory syndrome associated with COVID-19

Tripod checklist: Prediction model development and validation<sup>3</sup>

Section/Topic	Item	Checklist Item	Page	
<b>Title and abstract</b>				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	2
<b>Introduction</b>				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	4/5
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	5
<b>Methods</b>				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	5/6
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	5/6
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	5/6
	5b	D;V	Describe eligibility criteria for participants.	5
	5c	D;V	Give details of treatments received, if relevant.	NA
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	7
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	NA
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	8
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	NA
Sample size	8	D;V	Explain how the study size was arrived at.	5
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	9
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	12
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	12
	10c	V	For validation, describe how the predictions were calculated.	NA
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	NA
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	NA
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	NA
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	NA
<b>Results</b>				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	NA
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	NA
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	NA
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	NA
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	NA
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	NA
	15b	D	Explain how to use the prediction model.	NA
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	NA
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	NA
<b>Discussion</b>				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	NA
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	NA
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	NA
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	NA
<b>Other information</b>				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	NA
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	

## REFERENCES

1) Scottish Government. Management and recording of the long-term effects of COVID-19 (Long COVID). March 2020. Available from: <https://www.scimp.scot.nhs.uk/wp-content/uploads/CMO-Letter-09.03.21.pdf> (Accessed November 2021)

2) Public Health Scotland. Scottish Clinical Coding Standards. Number 27. <https://www.isdscotland.org/products-and-services/terminology-services/clinical-coding-guidelines/Docs/Scottish-clinical-coding-standards-Feb-2021-No-27.pdf> (Accessed November 2021)

3) Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine*. 2015;162(1):W1-73.