BMJ Open

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

Journal:	*BMJ Open*
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

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

Luke Daines*[1], Rachel H Mulholland*[1], Eleftheria Vasileiou[1], Vicky Hammersley[1], David Weatherill[1], Srinivasa Vittal Katikireddi[2], Steven Kerr[1], Emily Moore[3], Elisa Pesenti[4], Jennifer Quint[5], Syed Ahmar Shah[1], Ting Shi[1], Colin R Simpson[1,6], Chris Robertson[3,7], Aziz Sheikh[1]

*Contributed equally

1. Usher Institute, University of Edinburgh, Edinburgh, UK
2. MRC/CSO Social & Public Health Sciences Unit, University of Glasgow, Glasgow, UK
3. Public Health Scotland, Glasgow and Edinburgh, UK
4. Institute of Cell Biology, University of Edinburgh, Edinburgh, UK
5. Faculty of Medicine, National Heart & Lung Institute, Imperial College London, London, UK
6. School of Health, Wellington Faculty of Health, Victoria University of Wellington, Wellington, NZ
7. Department of Mathematics and Statistics, University of Strathclyde, Glasgow, UK

Corresponding author

Dr Luke Daines

Usher Institute, University of Edinburgh,

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

Email: luke.daines@ed.ac.uk

Word count: 3842 / 4000

ABSTRACT

Introduction

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

Methods and analysis

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

Ethics and dissemination

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

Word count: 295/300

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.

INTRODUCTION

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

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

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

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

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

METHODS AND ANALYSIS

Study design and population

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

Inclusion/exclusion criteria

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

Sample size calculation

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

Databases

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

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

Primary care data

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

Secondary care data

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

Laboratory data

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

Vaccination data

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

Telehealth data

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

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

Prescribing data

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

Mortality data

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

Other data

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

Determining an operational definition for long-COVID

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

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

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

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

Population characteristics

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

Socio-demographics

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

Geographical

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

Clinical characteristics

Using diagnostic codes from the QCOVID algorithm,[34] we will identify the following conditions: a) cardiovascular; b) diabetes (type 1 and type 2); c) respiratory d) cancer (blood cancer, chemotherapy, lung or oral cancer, marrow transplant, radiotherapy); e) neurological; f) other conditions, such as liver cirrhosis, osteoporotic fracture, rheumatoid arthritis, systemic lupus erythematosus, sickle cell disease, venous thromboembolism, solid organ transplant, renal failure (chronic kidney disease stages 3-5 with or without dialysis or transplant).[34]

Severity of acute COVD-19 illness

Admission to hospital, any requirement for treatment in the ICU, and death will be used to categorise the severity of COVID-19 infection. We will define a COVID-19 hospitalisation as a RT-PCR confirmed positive test for SARS-CoV-2 in the 28 days prior to admission, or admission with an ICD-10 code for COVID-19. A COVID-19 ICU admission will be defined as

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

Missing data

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

Statistical analyses

Developing an operational definition of long-COVID

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

Indicators of long-COVID

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

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

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

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

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

Matched analysis.

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

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

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

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

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

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

Cluster analysis

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

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

Sensitivity analyses

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

Deriving and validating a risk prediction model for long-COVID

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

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

Descriptive analysis

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

Outcome

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

Predictor variables

Predictors for the risk prediction model will consist of the patient characteristics, including information on socio-demographics, geographical, clinical comorbidities and severity of COVID-19 infection. These will be a mixture of continuous, binary and categorical variables. Continuous variables will be tested for linearity and for more flexible relationships (using smooth splines). Groupings of continuous variables will be explored if necessary.

Type of model

We intend to use a multivariable logistic regression model.

Selection of predictors

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

Model evaluation/performance

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

Model validation

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

Risk groups

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

Sensitivity analysis

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

Enhancing the prediction model using machine learning

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

Patient and public involvement

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

ETHICS AND DISSEMINATION

Ethical approval

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

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

Dissemination

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

Policy and Public Health

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

Academic

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

Community-based

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

AUTHORS CONTRIBUTIONS

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

FUNDING STATEMENT

This work was supported by the Chief Scientist Office, grant number COV/LTE/20/15. EAVE II is supported by a grant (MC_PC_19075) from the Medical Research Council; a grant (MC_PC_19004) from BREATHE–The Health Data Research Hub for Respiratory Health, funded through the UK. Research and Innovation Industrial Strategy Challenge Fund and delivered through Health Data Research UK; a grant from the Data and Connectivity National Core Study, led by Health Data Research UK in partnership with the Office for National Statistics and funded by UK Research and Innovation (grant ref MC_PC_20058); Public Health Scotland; and the Scottish Government Director General for Health and Social Care. SVK acknowledges funding from a NRS Senior Clinical Fellowship (SCAF/15/02), the Medical Research Council (MC_UU_00022/2) and the Scottish Government Chief Scientist Office (SPHSU17).

COMPETING INTERESTS STATEMENT

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

ACKNOWLEDGEMENTS

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

DATA AVAILABILITY

No additional data available.

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 (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 (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. Carfì 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
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 (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

socialandpublichealthsciencesunit/aboutus/covid19/inequalitiescovid/ (Accessed November 2021)
33. Generation Scotland. CovidLife. Available from: https://www.ed.ac.uk/generation-scotland/for-researchers/covidlife (Accessed November 2021)
34. Clift AK, Coupland CA, Keogh RH, et al. Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. *BMJ* 2020;371:m3731.
35. Nguyen AN, Truran D, Kemp M, et al. Computer-Assisted Diagnostic Coding: Effectiveness of an NLP-based approach using SNOMED CT to ICD-10 mappings. *AMIA Annu Symp Proc*. 2018:807-816.
36. Koleck TA, Dreisbach C, Bourne PE, et al. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J *Am Med Inform Assoc*. 2019:26(4):364-379
37. Gower JC. A General Coefficient of Similarity and Some of Its Properties. Biometrics 1971:1:857-871.
38. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987:20:53-65.
39. Dunn J.C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics* 1973:3(3)32-57, DOI: 10.1080/01969727308546046
40. 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.

FIGURE LEGENDS

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

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

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

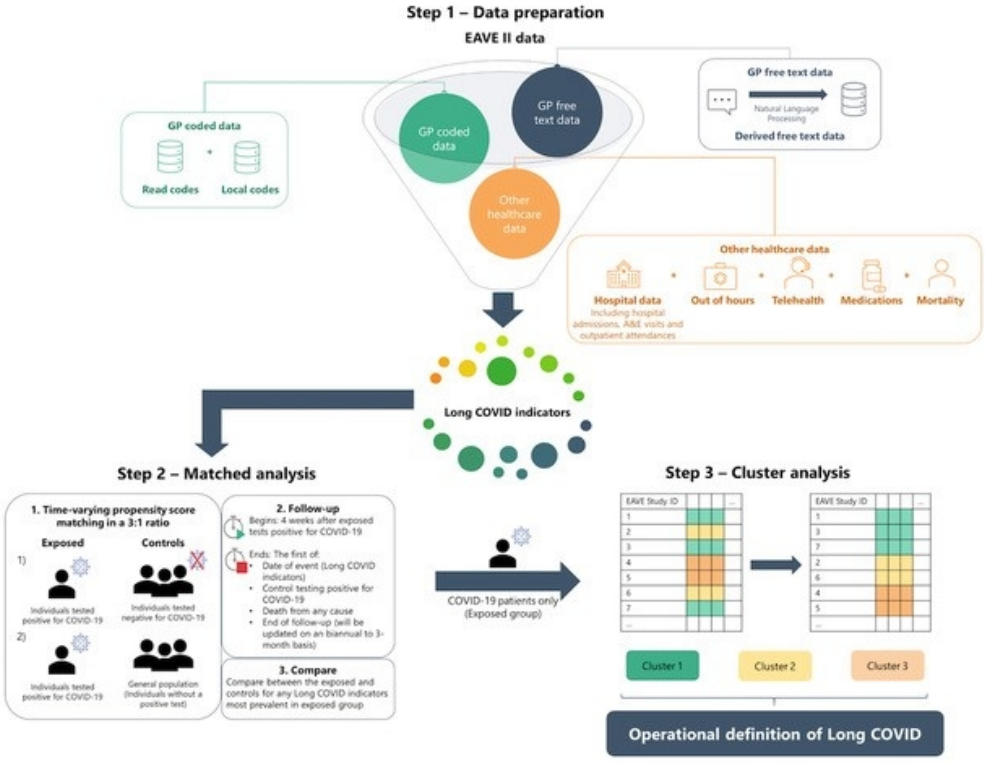

Figure 1: Schematic diagram of methods for developing an operational definition for Long COVID.

52x40mm (300 x 300 DPI)

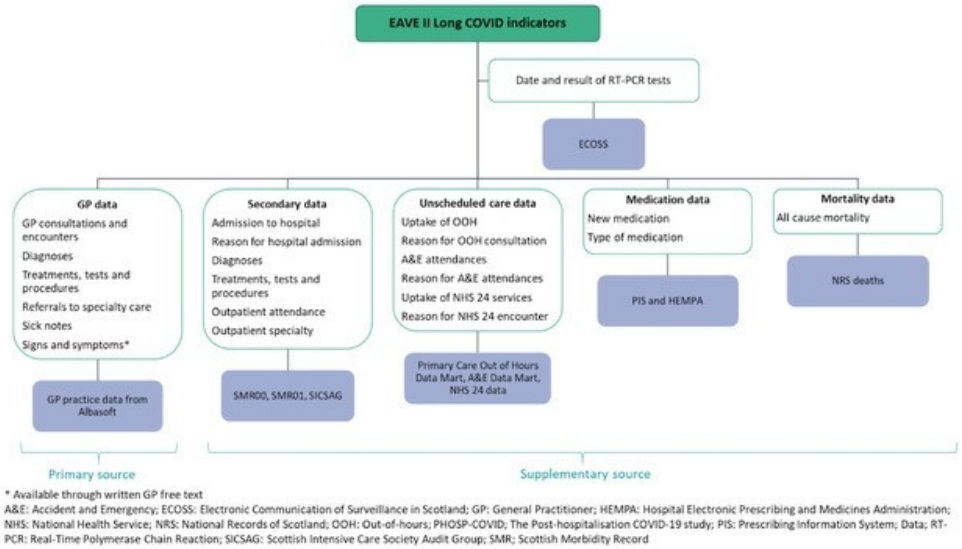

Figure 2: Data linkage diagram of long-COVID indicators and their data sources within the EAVE II cohort.

54x32mm (300 x 300 DPI)

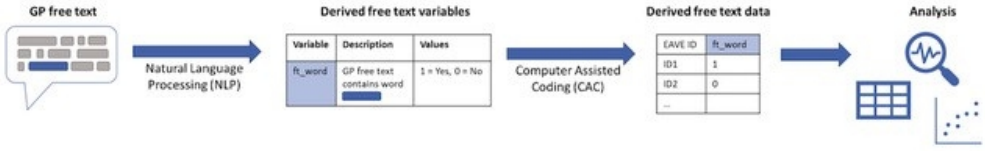

Figure 3: Schematic diagram of Natural Language Processing (NLP) and Computer Assisted Coding (CAC) on GP free text

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[1]

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[2]

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[3]

Section/Topic	Item		Checklist Item	Page
Title and abstract				
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
Introduction				
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
Methods				
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
Results				
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 the 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
Discussion				
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
Other information				
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.

BMJ Open

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

Journal:	*BMJ Open*
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
Primary Subject Heading:	Public health
Secondary Subject Heading:	Public health
Keywords:	PUBLIC HEALTH, COVID-19, Protocols & guidelines < HEALTH SERVICES ADMINISTRATION & MANAGEMENT

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

Luke Daines*[1], Rachel H Mulholland*[1], Eleftheria Vasileiou[1], Vicky Hammersley[1], David Weatherill[1], Srinivasa Vittal Katikireddi[2], Steven Kerr[1], Emily Moore[3], Elisa Pesenti[4], Jennifer Quint[5], Syed Ahmar Shah[1], Ting Shi[1], Colin R Simpson[1,6], Chris Robertson[3,7], Aziz Sheikh[1]

*Contributed equally

1. Usher Institute, University of Edinburgh, Edinburgh, UK
2. MRC/CSO Social & Public Health Sciences Unit, University of Glasgow, Glasgow, UK
3. Public Health Scotland, Glasgow and Edinburgh, UK
4. Institute of Cell Biology, University of Edinburgh, Edinburgh, UK
5. Faculty of Medicine, National Heart & Lung Institute, Imperial College London, London, UK
6. School of Health, Wellington Faculty of Health, Victoria University of Wellington, Wellington, NZ
7. Department of Mathematics and Statistics, University of Strathclyde, Glasgow, UK

Corresponding author

Dr Luke Daines

Usher Institute, University of Edinburgh,

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

Email: luke.daines@ed.ac.uk

Word count: 3944 / 4000

ABSTRACT

Introduction

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

Methods and analysis

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

Ethics and dissemination

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

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.

INTRODUCTION

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

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

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

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

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

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

METHODS AND ANALYSIS

Study design and population

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

Inclusion/exclusion criteria

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

Sample size calculation

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

Databases

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

Primary care data

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

Secondary care data

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

Laboratory data

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

Vaccination data

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

Telehealth data

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

Prescribing data

Prescription data relating to all medications prescribed and dispensed in the community in Scotland will be extracted from the Prescribing Information System (PIS).[26] These medications are coded using the British National Formulary (BNF) code lists.[31] For medication data within hospitals, Hospital Electronic Prescribing and Medicines Administration (HEPMA) which are available for five Health Boards will be used.[26]

Mortality data

Mortality data will be taken from death registry data within the National Records of Scotland. These records hold information included on the death certificate, including cause(s) of death which are recoded using ICD-10 codes.[26]

Other data

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

Determining an operational definition for long-COVID

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

1. Acute COVID-19 infection: individuals with signs and symptoms of COVID-19 for up to 4 weeks
2. Ongoing symptomatic COVID-19: individuals with signs and symptoms of COVID-19 from 4-12 weeks

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

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

Population characteristics

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

Socio-demographics

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

Geographical

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

Clinical characteristics

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

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

Severity of acute COVD-19 illness

Admission to hospital, any requirement for treatment in the ICU, and death will be used to categorise the severity of COVID-19 infection. We will define a COVID-19 hospitalisation as a RT-PCR confirmed positive test for SARS-CoV-2 in the 28 days prior to admission, or admission with an ICD-10 code for COVID-19. A COVID-19 ICU admission will be defined as a RT-PCR confirmed positive test for SARS-CoV-2 in the 28 days prior to ICU admission. A COVID-19 death will be defined as dying within 28 days of confirmed or probable COVID-19.

Missing data

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

Statistical analyses

Developing an operational definition of long-COVID

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

Indicators of long-COVID

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

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

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

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

Matched analysis.

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

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

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

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

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

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

Cluster analysis

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

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

Sensitivity analyses

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

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

Deriving and validating a risk prediction model for long-COVID

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

Descriptive analysis

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

Outcome

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

Predictor variables

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

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

Type of model

We intend to use a multivariable logistic regression model.

Selection of predictors

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

Model evaluation/performance

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

Model validation

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

Risk groups

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

Sensitivity analysis

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

Enhancing the prediction model using machine learning

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

Patient and public involvement

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

ETHICS AND DISSEMINATION

Ethical approval

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

Dissemination

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

Policy and Public Health

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

Academic

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

Community-based

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

AUTHORS CONTRIBUTIONS

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

FUNDING STATEMENT

This work was supported by the Chief Scientist Office, grant number COV/LTE/20/15. EAVE II is supported by a grant (MC_PC_19075) from the Medical Research Council; a grant (MC_PC_19004) from BREATHE–The Health Data Research Hub for Respiratory Health, funded through the UK. Research and Innovation Industrial Strategy Challenge Fund and delivered through Health Data Research UK; a grant from the Data and Connectivity National Core Study, led by Health Data Research UK in partnership with the Office for National Statistics and funded by UK Research and Innovation (grant ref MC_PC_20058); Public Health Scotland; and the Scottish Government Director General for Health and Social Care. SVK acknowledges funding from a NRS Senior Clinical Fellowship (SCAF/15/02), the Medical Research Council (MC_UU_00022/2) and the Scottish Government Chief Scientist Office (SPHSU17).

COMPETING INTERESTS STATEMENT

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

ACKNOWLEDGEMENTS

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

REFERENCES

1. World Health Organization. WHO Coronavirus Disease (COVID-19) Dashboard. Available from: https://covid19.who.int/ (Accessed November 2021)
2. Public Health Scotland. COVID-19 in Scotland Daily Dashboard. Available from: https://public.tableau.com/profile/phs.covid.19#!/vizhome/COVID-19DailyDashboard_15960160643010/Overview (Accessed November 2021)
3. 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 (Accessed November 2021)
4. 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)
5. Carfì A, Bernabei R, Landi F. Persistent Symptoms in Patients After Acute COVID-19. *JAMA* 2020;324(6):603-605.
6. del Rio C, Collins LF, Malani P. Long-term Health Consequences of COVID-19. *JAMA* 2020;324(17):1723-4.
7. 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
8. 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; 9(11):1275-1287 DOI: https://doi.org/10.1016/S2213-2600(21)00383-0
9. Ayoubkhani D, Khunti K, Nafilyan V, et al. Epidemiology of post-COVID syndrome following hospitalisation with coronavirus: a retrospective cohort study. medRxiv 2021.01.15.21249885; doi: https://doi.org/10.1101/2021.01.15.21249885
10. 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.
11. 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.
12. 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.
13. Rajkumar RP. COVID-19 and mental health: A review of the existing literature. *Asian J Psychiatr* 2020;52:102066.
14. 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)

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 (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)

28. Information Services Division. SMR Datasets. General Clinical Information. Available from: https://www.ndc.scot.nhs.uk/Data-Dictionary/SMR-Datasets/General-Clinical-Information/ (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/mrccsosocialandpublichealthsciencesunit/aboutus/covid19/inequalitiescovid/ (Accessed November 2021)
33. Generation Scotland. CovidLife. Available from: https://www.ed.ac.uk/generation-scotland/for-researchers/covidlife (Accessed November 2021)
34. Clift AK, Coupland CA, Keogh RH, et al. Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. *BMJ* 2020;371:m3731.
35. Nguyen AN, Truran D, Kemp M, et al. Computer-Assisted Diagnostic Coding: Effectiveness of an NLP-based approach using SNOMED CT to ICD-10 mappings. *AMIA Annu Symp Proc*. 2018:807-816.
36. Koleck TA, Dreisbach C, Bourne PE, et al. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J *Am Med Inform Assoc*. 2019:26(4):364-379
37. Gower JC. A General Coefficient of Similarity and Some of Its Properties. Biometrics 1971:1:857-871.
38. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987:20:53-65.
39. Dunn J.C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics* 1973:3(3)32-57, DOI: 10.1080/01969727308546046
40. 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.

FIGURE LEGENDS

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

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

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

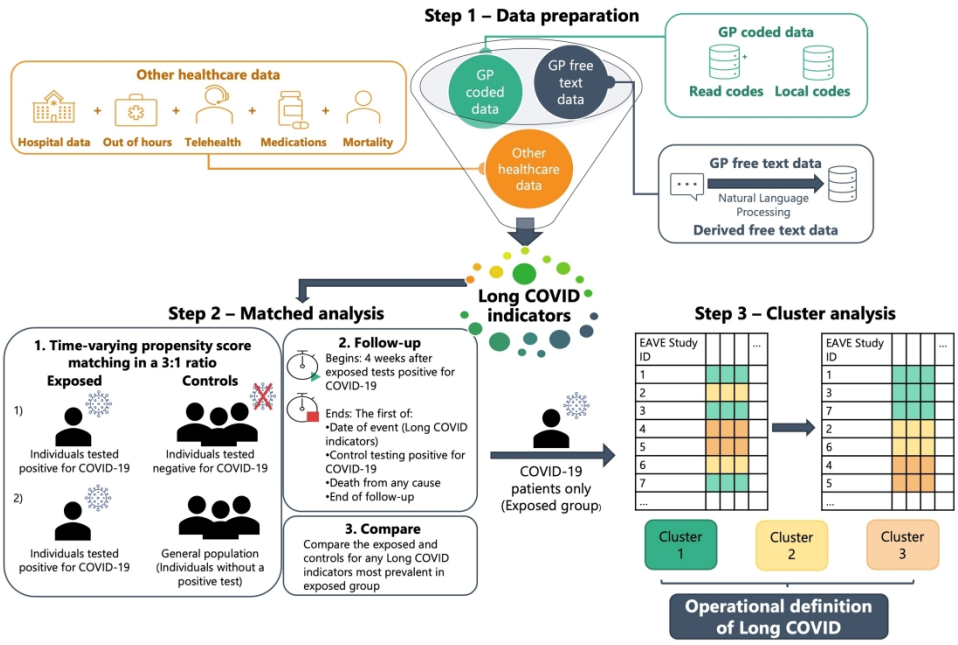

Figure 1: Schematic diagram of methods for developing an operational definition for Long COVID.

275x190mm (300 x 300 DPI)

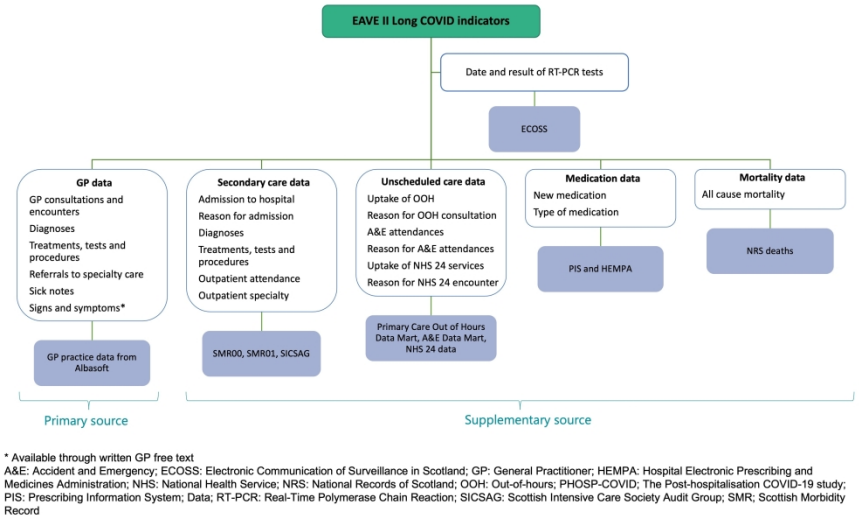

Figure 2: Data linkage diagram of long-COVID indicators and their data sources within the EAVE II cohort.

384x212mm (300 x 300 DPI)

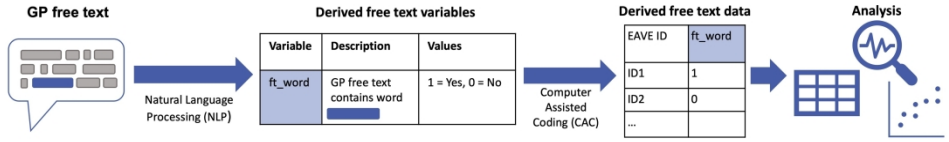

Figure 3: Schematic diagram of Natural Language Processing (NLP) and Computer Assisted Coding (CAC) on GP free text

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[1]

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[2]

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[3]

Section/Topic	Item		Checklist Item	Page
Title and abstract				
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
Introduction				
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
Methods				
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
Results				
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 the 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
Discussion				
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
Other information				
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.