Open Forum Infectious Diseases

MAJOR ARTICLE

An Elastic Net Regression Model for Identifying Long COVID Patients Using Health Administrative Data: A Population-Based Study

Mawuena Binka,[1,2] Braeden Klaver,[2] Georgine Cua,[1,2] Alyson W. Wong,[3,4] Chad Fibke,[2] Héctor A. Velásquez García,[1,2] Prince Adu,[1,2] Adeera Levin,[3] Sharmistha Mishra,[5,6] Beate Sander,[7,8] Hind Sbihi,[1,2] and Naveed Z. Janjua[1,2,9]

[1]School of Population and Public Health, University of British Columbia, Vancouver, British Columbia, Canada, [2]Data and Analytic Services, British Columbia Centre for Disease Control, Vancouver, British Columbia, Canada, [3]Department of Medicine, University of British Columbia, Vancouver, British Columbia, Canada, [4]Centre for Heart Lung Innovation, St. Paul's Hospital, University of British Columbia, Vancouver, British Columbia, Canada, [5]MAP Centre for Urban Health Solutions, St. Michael's Hospital, Toronto, Ontario, Canada, [6]Department of Medicine, University of Toronto, Toronto, Ontario, Canada, [7]Toronto Health Economics and Technology Assessment (THETA) Collaborative, University Health Network, Toronto, Ontario, Canada, [8]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada, and [9]Centre for Health Evaluation and Outcome Sciences, St Paul's Hospital, Vancouver, British Columbia V6Z IY6, Canada

Background. Long coronavirus disease (COVID) patients experience persistent symptoms after acute severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection. Healthcare utilization data could provide critical information on the disease burden of long COVID for service planning; however, not all patients are diagnosed or assigned long COVID diagnostic codes. We developed an algorithm to identify individuals with long COVID using population-level health administrative data from British Columbia (BC), Canada.

Methods. An elastic net penalized logistic regression model was developed to identify long COVID patients based on demographic characteristics, pre-existing conditions, COVID-19-related data, and all symptoms/conditions recorded >28–183 days after the COVID-19 symptom onset/reported (index) date of known long COVID patients (n = 2430) and a control group (n = 24 300), selected from all adult COVID-19 cases in BC with an index date on/before October 31, 2021 (n = 168 111). Known long COVID cases were diagnosed in a clinic and/or had the International Classification of Diseases, Tenth Revision, Canada (ICD-10-CA) code for "post COVID-19 condition" in their records.

Results. The algorithm retained known symptoms/conditions associated with long COVID, demonstrating high sensitivity (86%), specificity (86%), and area under the receiver operator curve (93%). It identified 25 220 (18%) long COVID patients among the remaining 141 381 adult COVID-19 cases, >10 times the number of known cases. Known and predicted long COVID patients had comparable demographic and health-related characteristics.

Conclusions. Our algorithm identified long COVID patients with a high level of accuracy. This large cohort of long COVID patients will serve as a platform for robust assessments on the clinical course of long COVID, and provide much needed concrete information for decision-making.

Keywords. long COVID; post-COVID-19 condition; post-acute COVID-19 syndrome; post-acute sequelae of COVID-19.

Approximately 20% of adult coronavirus disease 2019 (COVID-19) survivors develop long COVID, a syndrome characterized by a wide range of persistent symptoms and conditions affecting multiple body systems that emerge during or after the acute phase of COVID-19 illness [1–4]. Also known as post-COVID-19 condition, the full spectrum of symptoms and conditions that define long COVID is still under investigation, and the underlying mechanisms associated with this syndrome are poorly understood [3, 5]. Assessment for long COVID is done at 4 weeks (Centers for Disease Control and Prevention [CDC], United States; National Institute for Health and Care Excellence [NICE], United Kingdom [UK]) [6, 7] or 12 weeks (World Health Organization [WHO]; Government of Canada) [8, 9] after the initial COVID-19 infection. Commonly reported symptoms associated with this syndrome include malaise and fatigue, shortness of breath, myalgia, and brain fog/cognitive impairment [2, 3, 6, 8, 10, 11]. The broad range of symptoms and conditions experienced by long COVID patients has a profound impact on their quality of life, with 70%–86% needing to limit work schedules to accommodate this condition [2, 11].

Received 13 October 2022; editorial decision 18 November 2022; accepted 22 November 2022; published online 24 November 2022

Correspondence: Naveed Zafar Janjua, MBBS, MSc, DrPH, School of Population and Public Health, University of British Columbia, 655 West 12th Avenue, Vancouver, BC V5Z 4R4, Canada. E-mail: naveed.janjua@bccdc.ca, or Mawuena Binka, PhD, MPH, Data and Analytic Services, BC Centre for Disease Control, 655 West 12th Avenue, Vancouver, BC V5Z 4R4, Canada. E-mail: mawuena.binka@bccdc.ca.

Open Forum Infectious Diseases®

https://doi.org/10.1093/ofid/ofac640

Most of what is known about long COVID is based on data from clinical studies involving restricted subsets of the population [1, 2, 12]. However, given the large number of people projected to be living with long COVID worldwide, larger-scale population-level studies are needed to support exhaustive assessment of the factors associated with long COVID and to inform the clinical management of this complex condition. Gaps in the care continuum for persons living with long COVID could also be evaluated to provide concrete population-level data to ensure equitable access to care. These macro-level assessments of long COVID could be done using the International Classification of Diseases, Tenth Revision (ICD-10), code for "post COVID-19 condition, unspecified" in health administrative databases. However, preliminary data from the United States [13] and Canada suggest that this code for long COVID is not widely used, possibly due to the heterogeneous nature of this condition. Thus, algorithms using various patient characteristics are needed to identify long COVID patients in health administrative databases to facilitate the population-level assessment of this condition. Using linked health administrative datasets, we developed an algorithm to identify long COVID patients in a population-based cohort of COVID-19 cases in British Columbia (BC), Canada. Data from long COVID patients identified through this algorithm could support further robust assessments related to long COVID.

METHODS

Study Population

This study used the BC COVID-19 Cohort (BCC19C) [14], which includes data from all individuals who tested for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), and reported COVID-19 cases in BC. This information is integrated with data from various provincial registries/databases containing demographic, immunization, emergency room (ER) visit, medical visit, hospitalization, laboratory testing, prescription drug dispensation, chronic condition, and mortality data (Supplementary Table 1). Quantitative polymerase chain reaction (qPCR)–confirmed adult SARS-CoV-2 cases (≥18 years) with a COVID-19 index date (earliest date of symptom onset or reported date) on or before October 31, 2021, were eligible for this analysis (n = 168 111). The follow-up period for COVID-19-related symptom assessment began 28 days after the COVID-19 index date and lasted up to 183 days (6 months) afterward. The >28-day starting point was selected in accordance with CDC/NICE [6, 7] definitions of long COVID, although a sensitivity analysis was done with symptom assessment after 12 weeks (WHO/Canada definitions) [8, 9]. The shortest follow-up time for symptom assessment for this analysis was 3 months. Out-of-province COVID-19 cases and people who died any time during their follow-up period were excluded from this analysis.

Long COVID Status

Known long COVID patients in the BCC19C were either diagnosed in 1 of 4 Post-COVID-19 Recovery Clinics (PCRC) in BC [15] or identified with the Canadian ICD-10 (ICD-10-CA) code for "post COVID-19 condition" (U07.4) [16, 17] during ER visits or hospitalization. Patients seen in PCRCs were either hospitalized or treated as outpatients during their acute COVID-19 illness. Assessment for long COVID at PCRCs was done at baseline, 3 months, and 6 months after symptom onset using standardized validated questionnaires assessing patient-reported outcome measures (fatigue, Fatigue Severity Scale; cough, Cough Visual Analogue Scale; dyspnea, University of California San Diego shortness of breath questionnaire; anxiety, Generalized Anxiety Disorder 2-item; depression, Patient Health Questionnaire-2; post-traumatic stress disorder [PTSD], Primary Care Post Traumatic Stress Disorder Screen for DSM-5) [18–23] and quality of life (EuroQoL 5 dimensions visual analogue scale [EQ5D VAS]) [24]. PCRC-diagnosed long COVID patients had persistent symptoms for at least 12 weeks after acute symptom onset, and ≥1 abnormal patient-reported outcome measure score (cough, ≥30/100; dyspnea, ≥10/120; fatigue, ≥4/7; anxiety, ≥3/6; depression, ≥3/6; PTSD, ≥3/5) during their initial PCRC assessment.

In BC, a special ICD-9 code "C19" was introduced in March 2020 to denote medical visits for "services directly related to COVID-19" [25]. To minimize misclassification of long COVID cases within our control group, controls for the model were selected from the subset of the remaining eligible COVID-19 cases who did not visit a PCRC and had no occurrence of this C19 code after acute COVID-19 illness, that is, >14 days after their COVID-19 index date (Figure 1). This was based on the assumption that these individuals may not have been sick long enough to seek additional medical care outside of the hospital setting after the acute phase of their illness, and thus were less likely to be living with long COVID.

Machine Learning Approach

COVID-19 cases were divided into the development dataset, comprised of all known long COVID cases and controls, and the application dataset, which included all remaining COVID-19 patients (Supplementary Figure 1), to support the identification of additional long COVID patients using a machine learning approach: elastic net penalized logistic regression (*glmnet* package in R) [26–29]. The elastic net penalized logistic regression model [26–29] leverages the penalties applied in both ridge regression and lasso regression to perform variable selection in the presence of highly correlated independent variables, constructing a parsimonious model that is not overfitted to the development dataset. Our elastic net model was developed using the characteristics of all known long COVID patients (n = 2430) and controls (n = 24 300) in the cohort. Controls were randomly selected from the group of

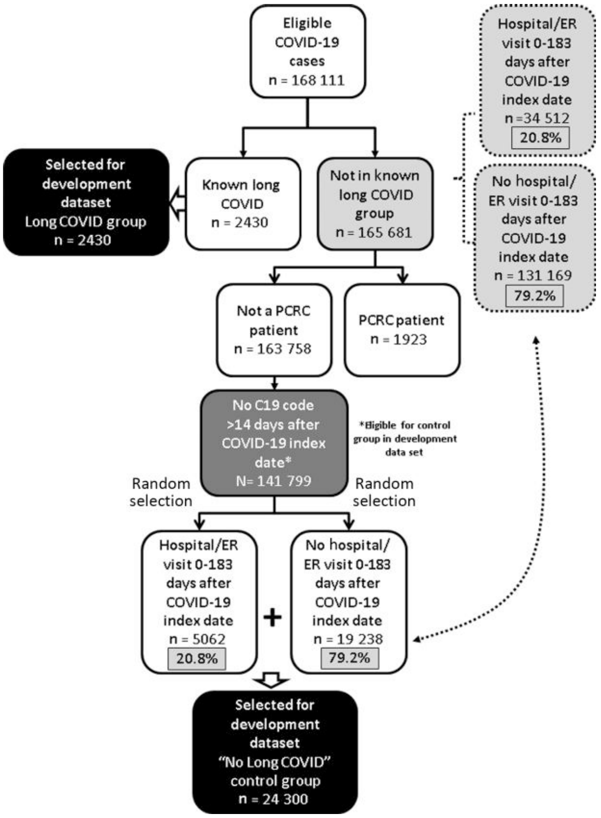
Eligible COVID-19 cases n = 168 111
Hospital/ER visit 0-183 days after COVID-19 index date n =34 512
20.8%
Selected for development dataset Long COVID group n = 2430
Known long COVID n = 2430
Not in known long COVID group n = 165 681
No hospital/ ER visit 0-183 days after COVID-19 index date n = 131 169
79.2%
Not a PCRC patient n = 163 758
PCRC patient n = 1923
No C19 code >14 days after COVID-19 index date* N= 141 799
*Eligible for control group in development data set
Random selection
Random selection
Hospital/ER visit 0-183 days after COVID-19 index date n = 5062
20.8%
+
No hospital/ ER visit 0-183 days after COVID-19 index date n = 19 238
79.2%
Selected for development dataset "No Long COVID" control group n = 24 300

Figure 1. Selection of the development dataset. COVID-19 cases were divided into the development dataset, comprised of all known long COVID cases and controls, and the application dataset, which included all remaining COVID-19 patients. To minimize misclassification of long COVID cases within our control group, controls for the model were selected from the subset of the remaining eligible COVID-19 cases who did not visit a PCRC and had no occurrence of the BC-specific "C19" ICD-9 code after acute COVID-19 illness (>14 days after their COVID-19 index date). Random sampling of controls was done to reflect the underlying distribution of the population pertaining to hospitalization/ER visits during the 6-month follow-up period (dotted sections). Abbreviations: COVID-19, coronavirus disease 2019; ER, emergency room; ICD-9, International Classification of Diseases, Ninth Revision; PCRC, Post-COVID-19 Recovery Clinic.

COVID-19 cases with decreased likelihood of living with long COVID, determined by the absence of the C19 code >14 days after the COVID-19 index date as described above (Figure 1). Random sampling of controls was done to reflect the underlying distribution of the population pertaining to hospitalization/ER visits during the 6-month follow-up period (Figure 1).

Variable Selection for the Model

Patient characteristics assessed within the elastic net regression model included (i) demographic and (ii) geographic variables; (iii) socioeconomic status, assessed using the Québec Index of Material and Social Deprivation [30]; (iv) pre-existing chronic conditions before the COVID-19 index date, including asthma, heart disease, and hypertension, as determined with ICD-9/10 diagnostic/intervention codes, billing codes, and prescription drug dispensations (Supplementary Table 2); (v) COVID-19-related data including SARS-CoV-2 variant, SARS-CoV-2 vaccination status, and hospitalization; and (vi) all relevant unique symptoms, conditions, and presenting complaints noted during medical/ER visits or hospitalization throughout the follow-up period (Supplementary Figure 2). Missing geographic region, socioeconomic status, age, and sex were denoted as "unknown," and, with exception of the 2 people with unknown ages, all other individuals with unknown sex, socioeconomic status, and geographic region were included in the models.

Modeling and Statistical Analyses

Two tuning parameters were used to optimize the elastic net regression model: alpha and lambda. Alpha determines the associated weight applied to either the ridge or the lasso regression penalty, which ranges from 0 (full ridge regression) to 1 (full lasso regression) [26]. Lambda is the applied combined penalty, which is the summation of either the squared or absolute regression coefficients for the ridge and lasso penalties, respectively [26]. We applied 11 different values of alpha (0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0) to the penalized elastic net model using the *glmnet* package (version 3.0-2) [29]. Lambda determinations to optimize the area under the receiver operator characteristic curve (AUROC) were done with 10-fold cross-validation in the development data set. A predictive probability threshold was determined using the *ROCR* package (version 1.0-7) that maximized both sensitivity and specificity equally for the model. The best model was selected based on specificity, sensitivity, and AUROC. All analyses were done with R statistical software (version 3.6.2) [31]. The development data set was also split 80% for training and 20% for testing (*caret* package [32], version 6.0-85) for comparison (Supplementary Figure 1).

Sensitivity Analyses

Additional analyses were conducted to assess the effect of choices made on model output: First, we examined the effect of solely using PCRC-confirmed known long COVID patients in the "known long COVID" group. Second, to reduce the risk of misclassifying long COVID patients as controls, we restricted the pool of possible controls to individuals who were not PCRC patients and had no record of a C19 code >14 days after their COVID-19 index date in the base case analysis. In a sensitivity analysis, all individuals who were not PCRC patients and had no record of a C19 code throughout the 6-month follow-up period were excluded from the control group. Third, to determine the impact of the >28–183-day symptom evaluation period, additional models were run incorporating only symptoms/conditions recorded from (i) >84–183 days and (ii) 0–183 days post–COVID-19 index date. Finally, we evaluated the impact of excluding hospitalization-related variables (hospitalization, intensive care unit [ICU], life support) from the model.

Ethical approval for this study was provided by the University of British Columbia Behavioral Research Ethics Board (No: H20-02097).

RESULTS

Characteristics of Study Cohort Used for Model Development

Long COVID patients in the development data set were more likely to be female (53.9%), aged between 50 and 59 years old (23.2%), and hospitalized for COVID-19-related reasons (44.0%) (Table 1). Prevalent conditions within the long COVID group included diabetes mellitus (22.8%), hypertension (32.7%), depression (43.4%), and mood and anxiety disorders (51.4%). In contrast, a relatively smaller proportion of controls belonged to the 50–59 age group (13.1%), were female (48.4%), or hospitalized (1.8%) (Table 1). Baseline characteristics for long COVID patients differed by whether patients were hospitalized and/or visited the ER with COVID-19-related illness, such that those who received COVID-19-related care in the hospital setting were more likely to be older, with relatively higher prevalence of most comorbidities assessed.

Model Selection

Model robustness in the development data set was demonstrated by high sensitivity (≥80%), specificity (≥85%), and AUROC (≥92%) across the range of alphas assessed (Supplementary Figure 3, Supplementary Table 3). Models were also similar in variable composition (Supplementary Figure 4). Given the small variation in sensitivity, specificity, and AUROC, the alpha = 0.5 model, which leverages the best features of both ridge and lasso regression, was selected for application (alpha = 0.5, sensitivity = 86%, specificity = 86%, AUROC = 93%, probability threshold = 0.391). Long COVID–related variables selected by the optimal model encompassed each of the categories considered for inclusion. Top symptoms included shortness of breath, malaise/fatigue, and chest pain (Figure 2; Supplementary Table 4), while mood and anxiety disorders and heart disease were among pre-existing conditions in the model. Factors associated with SARS-CoV-2 infection retained in the model included variant of concern, vaccination status at COVID-19 index date, and hospitalization. Older age (40–49, 50–59, and 60–69 years) was a positive predictor of long COVID (Figure 2; Supplementary Table 4), while negative predictors included male sex and receiving at least 1 dose of a 2-dose COVID-19 vaccine (Supplementary Table 5). The top 50 positive and negative predictors included in the best performing model are shown in Supplementary Tables 4 and 5, respectively.

Characteristics of Known and Predicted Long COVID Cases

A total of 25 220 model-predicted long COVID cases were identified following model application, representing 18% of the application dataset. Table 2 shows the profiles of model-classified long COVID patients and the known long COVID patients whose characteristics were used to develop the model. Known and predicted long COVID patients were comparable in sex and age distribution, although model-predicted long COVID cases had a slightly larger proportion of females and of people aged 40–59 years (known/predicted rate ratio <1.0). Pre-existing conditions were also comparably distributed between both groups, with rate ratios nearing or equaling 1 for most conditions assessed (Table 2). These results were similar across models (Supplementary Figure 5). Notably, predicted long COVID cases had lower rates of severe disease, demonstrated by smaller proportions of hospitalizations, ICU admissions, and people on life support (rate ratios ≥2.1) (Table 2).

Sensitivity Analyses

Supplementary Table 6 summarizes data from the alpha = 0.5 models from the remaining sensitivity analyses. Selecting only PCRC-confirmed known long COVID patients improved model performance and increased the number of predicted cases. However, considering the measured use of this ICD-10-CA code in our health administrative datasets and the source-specific differences in known long COVID patient profiles (Supplementary Table 7), we opted to include known long COVID patients from both data sources. Restricting the pool of controls to COVID-19 cases without the C19 code recorded throughout their follow-up period resulted in modest improvements in model performance and an increased number of predicted long COVID cases. Nevertheless, given the resultant small shift in predicted long COVID patient characteristics (<1 percentage point), and to avoid eliminating recovered COVID-19 cases with severe acute illnesses from contention as possible controls, we opted for a 14-day cutoff for this code. Furthermore, adding all symptoms/conditions following COVID-19 diagnosis resulted in marginal improvements in sensitivity and specificity, decreasing the number of predicted cases. In addition, starting symptom assessment at >84 days (WHO definition) decreased model sensitivity and specificity, increasing the number of predicted long COVID cases. However, we decided to maintain symptom assessment at >28 days as long COVID is typically assessed after acute illness. Finally, removing variables related to hospitalization worsened model performance and increased the number of predicted cases. However, we opted to retain hospitalization variables in the model given the possible association between disease severity and long COVID.

DISCUSSION

In this study, we developed a computable phenotype model to identify people living with long COVID within a large

Table 1. Characteristics of Study Cohort Used for Model Development

	Known Long COVID			No Long COVID Controls		
	All (n = 2430) No. (%)	Hospitalization/ER Visit During Acute COVID-19 Illness[a] (n = 1498) No. (%)	No Hospitalization/ER Visit During Acute COVID-19 Illness[a] (n = 932) No. (%)	All (n = 24 300) No. (%)	Hospitalization/ER Visit During Acute COVID-19 Illness[a] (n = 2383) No. (%)	No Hospitalization/ER Visit During Acute COVID-19 Illness[a] (n = 21 917) No. (%)
Demographics	…	…	…	…	…	…
Age group	…	…	…	…	…	…
18–29 y	215 (8.8)	86 (5.7)	129 (13.8)	7857 (32.3)	468 (19.6)	7389 (33.7)
30–39 y	361 (14.9)	161 (10.7)	200 (21.5)	5635 (23.2)	511 (21.4)	5124 (23.4)
40–49 y	529 (21.8)	281 (18.8)	248 (26.6)	4313 (17.7)	489 (20.5)	3824 (17.4)
50–59 y	564 (23.2)	354 (23.6)	210 (22.5)	3195 (13.1)	388 (16.3)	2807 (12.8)
60–69 y	435 (17.9)	346 (23.1)	89 (9.5)	1948 (8.0)	276 (11.6)	1672 (7.6)
70–79 y	234 (9.6)	195 (13.0)	39 (4.2)	890 (3.7)	160 (6.7)	730 (3.3)
80+ y	92 (3.8)	75 (5.0)	17 (1.8)	462 (1.9)	91 (3.8)	371 (1.7)
Sex	…	…	…	…	…	…
Female	1310 (53.9)	709 (47.3)	601 (64.5)	11 764 (48.4)	1178 (49.4)	10 586 (48.3)
Male	1119 (46.0)	788 (52.6)	331 (35.5)	12 503 (51.5)	1200 (50.4)	11 303 (51.6)
Unknown	<5	<5	<5	33 (0.1)	<5	28 (0.1)
Pre-existing conditions[b]	…	…	…	…	…	…
Asthma	777 (32.0)	494 (33.0)	283 (30.4)	4320 (17.8)	615 (25.8)	3705 (16.9)
Chronic obstructive pulmonary disease	146 (6.0)	119 (7.9)	27 (2.9)	508 (2.1)	114 (4.8)	394 (1.8)
Chronic kidney disease	197 (8.1)	159 (10.6)	38 (4.1)	580 (2.4)	112 (4.7)	468 (2.1)
Diabetes mellitus	553 (22.8)	445 (29.7)	108 (11.6)	2047 (8.4)	352 (14.8)	1695 (7.7)
Heart disease	581 (23.9)	421 (28.1)	160 (17.2)	2309 (9.5)	424 (17.8)	1885 (8.6)
Hypertension	795 (32.7)	609 (40.7)	186 (20.0)	3569 (14.7)	535 (22.5)	3034 (13.8)
Chronic vascular disease	258 (10.6)	185 (12.3)	73 (7.8)	1182 (4.9)	224 (9.4)	958 (4.4)
Cancer	259 (10.7)	189 (12.6)	70 (7.5)	1213 (5.0)	186 (7.8)	1027 (4.7)
Depression	1055 (43.4)	633 (42.3)	422 (45.3)	6191 (25.5)	874 (36.7)	5317 (24.3)
Mood and anxiety disorders	1250 (51.4)	747 (49.9)	503 (54.0)	7656 (31.5)	1040 (43.6)	6616 (30.2)
COVID-19 related	…	…	…	…	…	…
Variant of concern	…	…	…	…	…	…
Alpha	291 (12.0)	199 (13.3)	92 (9.9)	2701 (11.1)	254 (10.7)	2447 (11.2)
Beta	0 (0.0)	0 (0.0)	0 (0.0)	12 (0.0)	<5	11 (0.1)
Delta	275 (11.3)	199 (13.3)	76 (8.2)	4710 (19.4)	585 (24.5)	4125 (18.8)
Gamma	284 (11.7)	232 (15.5)	52 (5.6)	1421 (5.8)	184 (7.7)	1237 (5.6)
Not a variant of concern	1580 (65.0)	868 (57.9)	712 (76.4)	15 456 (63.6)	1359 (57.0)	14 097 (64.3)
Hospitalized	1069 (44.0)	1069 (71.4)	0 (0.0)	443 (1.8)	443 (18.6)	0 (0.0)
ICU	541 (22.3)	541 (36.1)	0 (0.0)	72 (0.3)	72 (3.0)	0 (0.0)
Life support used	…	…	…	…	…	…
Mechanical ventilation	183 (7.5)	183 (12.2)	0 (0.0)	13 (0.1)	13 (0.5)	0 (0.0)
Oxygen	143 (5.9)	143 (9.5)	0 (0.0)	21 (0.1)	21 (0.9)	0 (0.0)
Other support used	276 (11.4)	276 (18.4)	0 (0.0)	28 (0.1)	28 (1.2)	0 (0.0)
No support used	1828 (75.2)	896 (59.8)	932 (100.0)	24 238 (99.7)	2321 (97.4)	21 917 (100.0)

Abbreviations: COVID, coronavirus disease; ER, emergency room; ICU, intensive care unit.

[a] 0–14 days after COVID-19 index date.

[b] Before COVID-19 index date.

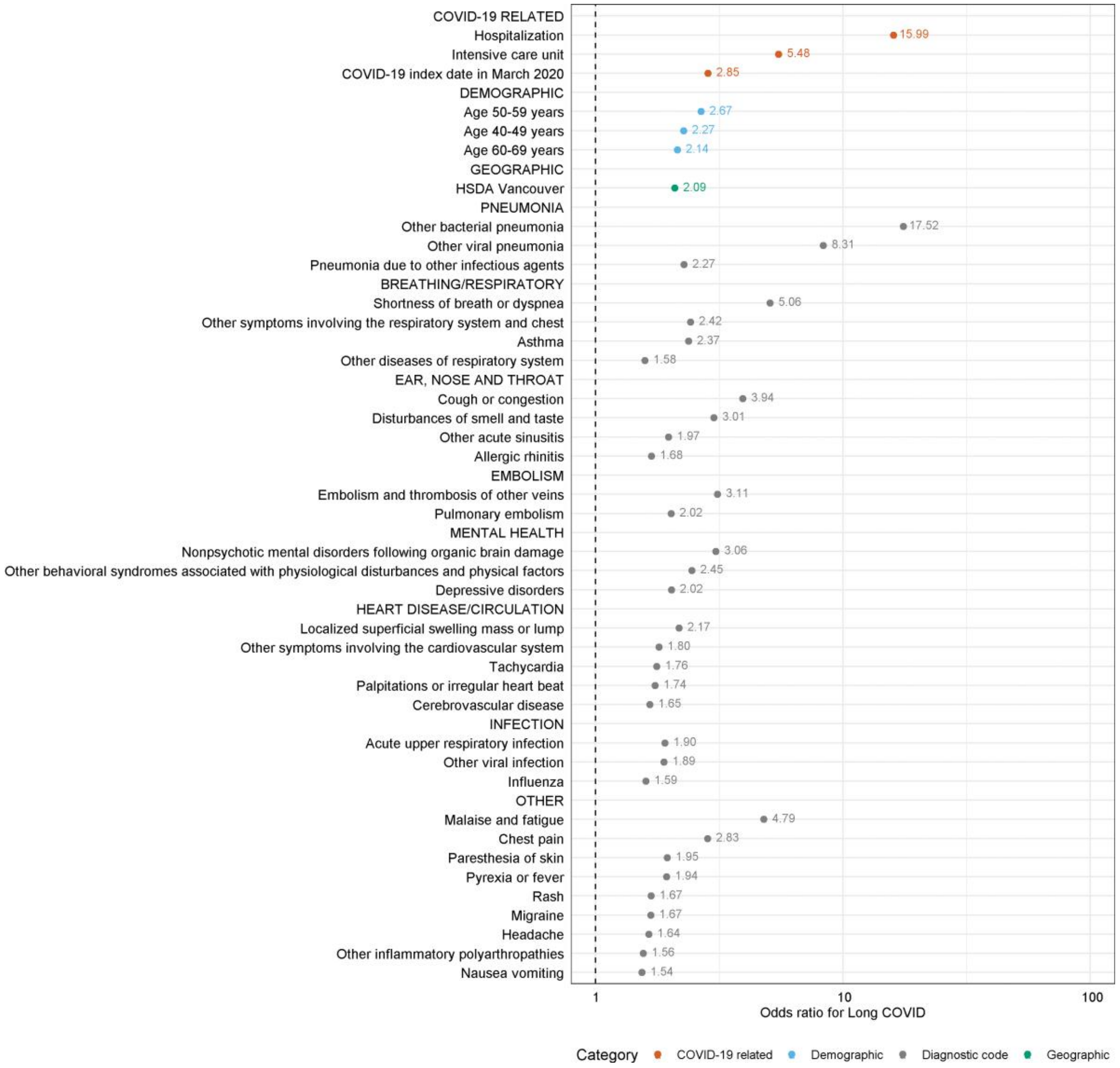

Figure 2. Top 40 positive predictors of long COVID status in the optimal model (alpha = 0.5). Determined in a cohort of known long COVID patients (n = 2430) and controls (n = 24 300). Abbreviation: COVID-19, coronavirus disease 2019.

population-based cohort of all reported COVID-19 cases in British Columbia, Canada. This elastic net regression model was trained on the characteristics of known long COVID patients within the cohort, including demographic variables, pre-existing conditions, COVID-19-related variables, and all unique symptoms/conditions/presenting complaints recorded at or after the COVID-19 index date. The optimal model had high sensitivity (86%), specificity (86%), and AUROC (93%), classifying 25 220 individuals as long COVID cases out of a possible 141 381 COVID-19 patients.

Variables identified by our model as being predictive of long COVID encompassed the major categories of factors commonly linked to the syndrome in several clinical studies [1, 2, 10, 12, 33]. Many of the top-ranking factors in the model were symptoms consistently reported by long COVID patients, such as shortness of breath/dyspnea and malaise and fatigue [1–3, 10, 12], indicating the robustness of our approach. Model stability was demonstrated with multiple sensitivity analyses showing minimal changes to the AUROC, sensitivity, and specificity, with changes in the symptom/condition assessment period and the exclusion of hospitalization-related variables. It was encouraging to note that the final model also considered demographic factors, socioeconomic status, and comorbid conditions alongside SARS-CoV-2/COVID-19-related factors when making determinations related to long COVID status. The direction of association between these variables and long COVID within the model was also congruent with known trends, such that the aforementioned symptoms and older age were predictive of long COVID [10], while male sex and vaccination were protective [12], further bolstering the validity of our model. Of

Table 2. Characteristics of Known and Predicted Long COVID Patients

	Known Long COVID (n = 2430)	Predicted Long COVID (n = 25 220)	
	No. (%)	No. (%)	Rate Ratio[b].
Demographics	…	…	…
Age group	…	…	…
18–29 y	215 (8.8)	1964 (7.8)	1.1
30–39 y	361 (14.9)	3815 (15.1)	1.0
40–49 y	529 (21.8)	6173 (24.5)	0.9
50–59 y	564 (23.2)	6413 (25.4)	0.9
60–69 y	435 (17.9)	3877 (15.4)	1.2
70–79 y	234 (9.6)	1929 (7.6)	1.3
80+ y	92 (3.8)	1049 (4.2)	0.9
Sex	…	…	…
Female	1310 (53.9)	15 860 (62.9)	0.9
Male	1119 (46.0)	9319 (37.0)	1.2
Unknown	<5	<5	…
Pre-existing conditions[b]	…	…	…
Asthma	777 (32.0)	7659 (30.4)	1.1
Chronic obstructive pulmonary disease	146 (6.0)	1502 (6.0)	1.0
Chronic kidney disease	197 (8.1)	1990 (7.9)	1.0
Diabetes mellitus	553 (22.8)	5014 (19.9)	1.1
Heart disease	581 (23.9)	5964 (23.6)	1.0
Hypertension	795 (32.7)	7903 (31.3)	1.0
Chronic vascular disease	258 (10.6)	2857 (11.3)	0.9
Cancer	259 (10.7)	2771 (11.0)	1.0
Depression	1055 (43.4)	12 954 (51.4)	0.8
Mood and anxiety disorders	1250 (51.4)	15 304 (60.7)	0.8
COVID-19 related	…	…	…
Variant of concern	…	…	…
Alpha	291 (12.0)	2757 (10.9)	1.1
Beta	0 (0.0)	22 (0.1)	0.0
Delta	275 (11.3)	2794 (11.1)	1.0
Gamma	284 (11.7)	2139 (8.5)	1.4
Not a variant of concern	1580 (65.0)	17 508 (69.4)	0.9
Hospitalized	1069 (44.0)	5316 (21.1)	2.1
ICU	541 (22.3)	1670 (6.6)	3.4
Life support used	…	…	…
Mechanical ventilation	183 (7.5)	393 (1.6)	4.7
Oxygen	143 (5.9)	499 (2.0)	3.0
Other support used	276 (11.4)	833 (3.3)	3.5
No support used	1828 (75.2)	23 495 (93.2)	0.8

Abbreviations: COVID, coronavirus disease; ER, emergency room, ICU, intensive care unit.

[a] Before COVID-19 index date.

[b] Rate ratio comparing prevalence among known vs predicted long COVID patients (known/predicted).

note, substance use and extreme material deprivation were among the negative predictive factors within the model, which speaks to the risk profiles of known long COVID patients in our cohort and may be reflective of systemic barriers to accessing healthcare among people with limited resources. The use of diagnostic codes for "other bacterial pneumonia" and "other viral pneumonia" appeared to mirror the COVID-19 waves in BC (data not shown), suggesting that clinicians may have been using this code as part of the clinical diagnosis process for people with severe COVID-19 illness, leading to strong associations with long COVID. Nevertheless, despite the large odds ratios for these conditions, only a fraction of model-predicted long COVID cases had at least 1 occurrence of these diagnostic codes (data not shown).

Our model identified 18% of our remaining population-based cohort of adult COVID-19 patients as probable long COVID cases. This is in line with published prevalence estimates of up to 20% within this population [1]. Together, profiles of known and predicted long COVID patients in our cohort paint a clearer picture of long COVID as a condition more likely to affect females and middle-aged people with pre-existing comorbidities such as mood and anxiety disorders, asthma, cardiovascular disease, and diabetes mellitus. These findings are in keeping with data from the United States, where chronic conditions such as diabetes were prevalent among hospitalized long COVID patients, who were also mostly female and middle-aged [12], and with data from the United Kingdom, where nonhospitalized long COVID patients also had an increased likelihood of being female and having pre-existing conditions such as asthma, anxiety, and depression, despite being relatively younger than people in our cohort [34].

This study is strengthened by the use of machine learning coupled with a hands-off approach to variable selection, which allowed for an objective assessment of long COVID–associated variables. Additionally, data fed into our model were derived from administrative datasets, which allowed for more uniform symptom assessment. Furthermore, our cohort was based on data from the entire population of BC, as opposed to selected clinical cohorts that select for patients with more severe disease [1, 12], making our study less prone to selection bias and our findings more applicable to the entire population. Although not accounted for in a similar study [12], the potential misclassification presented by the absence of confirmed "no long COVID" cases was addressed to the best of our ability using a BC-specific code for reporting COVID-19-related medical care. We lacked access to data pertaining to race/ethnicity and immigration status, which are associated with the comorbid conditions and socioeconomic factors selected by our model and could in turn be associated with long COVID. Future studies are planned to address this knowledge gap. We also plan to develop and validate algorithms to for identifying people with pre-existing autoimmune disorders and other chronic illnesses to provide additional dimension to the classification of long COVID patients by our model. Despite these limitations, the likelihood that model-predicted long COVID cases are true cases is increased by model performance, which was comparable to that of a similar model developed in the United States

[12], and by the strong similarity between the risk profiles of known and predicted long COVID cases. Given the heterogeneous nature of long COVID, misdiagnosis of some known long COVID cases was possible. At the very least, model-predicted long COVID cases represent individuals who should be assessed for long COVID and linkage with care, thus fulfilling the objective of providing critical information for service planning at the population level.

In this study, we developed an algorithm to identify long COVID patients at the population level using healthcare administrative datasets. Our model was developed with the best available data for the comprehensive assessment of symptoms and conditions associated with long COVID in BC. This model is reproducible and can be expanded to include additional datasets, such as immigration and ethnicity data, once they become available. Population-level data derived from the cohort of long COVID patients identified using this method are much needed to inform the clinical management of this condition and to provide concrete information for policy-making for the benefit of people living with long COVID.

Supplementary Data

Supplementary materials are available at *Open Forum Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Acknowledgments

We acknowledge the assistance of the Provincial Health Services Authority, BC Ministry of Health and Regional Health Authority staff involved in data access, procurement, and management. We gratefully acknowledge the residents of British Columbia whose data are integrated in the British Columbia COVID-19 Cohort (BCC19C), for whom this work is intended to benefit.

Financial support. The British Columbia COVID-19 Cohort was established and is maintained through operational support from Data Analytics, Reporting and Evaluation (DARE), and BC Centre for Disease Control (BCCDC) at the Provincial Health Services Authority. This work was supported by the Canadian Institutes of Health Research (grant number GA4-177757), the British Columbia Centre for Disease Control , and a Canada Research Chair in Economics of Infectious Diseases held by Beate Sander (CRC-950-232429). The sponsoring agencies had no role in the study design, data collection, data analysis, or interpretation.

Disclaimer. All inferences, opinions, and conclusions drawn in this publication are those of the authors and do not necessarily reflect the opinions or policies of the Data Steward(s).

Potential conflicts of interest. N.J.Z. has advised and spoken for AbbVie, not related to this work. All remaining authors declare no competing interests. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed.

Author contributions. M.B. and N.Z.J. designed the study. M.B. and B.K. directly accessed and verified the underlying data reported in this manuscript. M.B. and B.K. performed data analysis and visualization. M.B., N.Z.J., and B.K. contributed to data interpretation. G.C., A.W.W., C.F., H.V.G., P.A., R.B., A.L., S.M., B.S., and H.S. supported data interpretation and provided critical review of this work. All authors contributed to this manuscript and approved the final submission. M.B. and N.Z.J. were responsible for the decision to submit the manuscript for publication.

Statement of data availability. The data that support the findings are from the British Columbia COVID-19 Cohort and are not publicly available. Access to the data may be provided through the British Columbia Centre for Disease Control Institutional Data Access process to researchers who meet the criteria for accessing confidential data.

Patient consent. This study was performed using de-identified data routinely collected as part of public health surveillance and/or routine healthcare encounters. This study was reviewed and approved by the University of British Columbia Behavioral Research Ethics Board (No: H20-02097).

References

1. Bull-Otterson L. Post–COVID conditions among adult COVID-19 survivors aged 18–64 and ≥ 65 years—United States, March 2020–November 2021. MMWR Morb Mortal Week Rep **2022;** 71:713–7.
2. Davis HE, Assaf GS, McCorkell L, et al. Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. eClinicalMedicine **2021;** 38:101019.
3. Crook H, Raza S, Nowell J, Young M, Edison P. Long COVID—mechanisms, risk factors, and management. BMJ **2021;** 374:n1648.
4. World Health Organization. Coronavirus disease (COVID-19): post COVID-19 condition. Available at: https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-(covid-19)-post-covid-19-condition?gclid=EAIaIQobChMI6c3f07m4-QIV1cLCBB0oVA1bEAAYASAAEgIMNfD_BwE. Accessed August 8, **2022**.
5. Munblit D, O'Hara ME, Akrami A, Perego E, Olliaro P, Needham DM. Long COVID: aiming for a consensus. Lancet Respir Med **2022;** 10:632–4.
6. Centers for Disease Control and Prevention. Long COVID or post-COVID conditions. Available at: https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html. Accessed August 4, **2022**.
7. National Institute for Health and Care Excellence, Scottish Intercollegiate Guidelines Network, Royal College of General Practitioners. COVID-19 rapid guideline: managing the long-term effects of COVID-19 v1.14. Available at: https://www.nice.org.uk/guidance/ng188/resources/covid19-rapid-guideline-managing-the-longterm-effects-of-covid19-pdf-51035515742. Accessed August 4, **2022**.
8. World Health Organization. A clinical case definition of post COVID-19 condition by a Delphi consensus, 6 October 2021. Available at: https://www.who.int/publications/i/item/WHO-2019-nCoV-Post_COVID-19_condition-Clinical_case_definition-2021.1. Accessed August 4, **2022**.
9. Government of Canada. Post COVID-19 condition (long COVID). Available at: https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection/symptoms/post-covid-19-condition.html. Accessed August 4, **2022**.
10. O'Keefe JB, Minton HC, Morrow M, et al. Postacute sequelae of SARS-CoV-2 infection and impact on quality of life 1–6 months after illness and association with initial symptom severity. Open Forum Infect Dis **2021;** 8:XXX–XX.
11. Viral Neuro Exploration, COVID Long-Haulers Support Group Canada, Neurological Health Charities Canada. Report on Pan-Canadian Long Covid Impact Survey. Available at: https://imgix.cosmicjs.com/d8d3d3b0-c936-11eb-ba89-e7f98c8c358b-FINAL—Report-on-Long-Covid-Impact-Survey—June-8-2021.pdf. Accessed Dec 12, 2022.
12. Pfaff ER, Girvin AT, Bennett TD, et al. Identifying who has long COVID in the USA: a machine learning approach using N3C data. Lancet Digital Health **2022;** 4:e532–41.
13. Pfaff ER, Madlock-Brown C, Baratta JM, et al. Coding long COVID: characterizing a new disease through an ICD-10 lens. medRxiv 2022.04.18.22273968 [Preprint]. September 2, **2022**. Available at: https://doi.org/10.1101/2022.04.18.22273968.
14. UBC Centre for Disease Control. BC COVID-19 cohort. Available at: https://a4ph.med.ubc.ca/projects-and-initiatives/bc-covid-19-cohort/. Accessed June 6, **2022**.
15. Providence Health Care. Post-COVID-19 recovery clinic. Available at: https://www.providencehealthcare.org/covidrecoveryclinic. Accessed June 6, **2022**.
16. Canadian Institute for Health Information. CIHI portal release notes: release 16.1. Available at: https://www.cihi.ca/en/bulletin/cihi-portal-release-notes-release-161. Accessed June 6, **2022**.
17. Canadian Institute for Health Information. COVID-19: locating the ICD-10-CA/CCI code. Available at: https://www.cihi.ca/sites/default/files/document/covid-19-locating-icd-10-ca-cci-code-jobaid-en.pdf. Accessed June 6, **2022**.
18. Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. Med Care **2003;** 41:1284–92.
19. Kroenke K, Spitzer RL, Williams JB, Monahan PO, Lowe B. Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. Ann Intern Med **2007;** 146:317–25.

20. Kupferberg DH, Kaplan RM, Slymen DJ, Ries AL. Minimal clinically important difference for the UCSD Shortness of Breath Questionnaire. J Cardiopulm Rehabil **2005;** 25:370–7.
21. Nguyen A M, Bacci ED, Vernon M, et al. Validation of a visual analog scale for assessing cough severity in patients with chronic cough. Ther Adv Respir Dis **2021;** 15:17534666211049743.
22. Prins A, Bovin MJ, Smolenski DJ, et al. The primary care PTSD screen for DSM-5 (PC-PTSD-5): development and evaluation within a veteran primary care sample. J Gen Intern Med **2016;** 31:1206–11.
23. Valko PO, Bassetti CL, Bloch KE, Held U, Baumann CR. Validation of the fatigue severity scale in a Swiss cohort. Sleep **2008;** 31:1601–7.
24. EuroQol. EuroQoL instruments. Available at: https://euroqol.org/eq-5d-instruments/. Accessed February 10, **2021**.
25. BC Family Doctors. Billing in the time of COVID-19. Available at: https://bcfamilydocs.ca/covid19/#collapse-Covid-Billing-18. Accessed June 6, **2022**.
26. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Stat Methodol) **2005;** 67:301–20.
27. Hastie T, Qian J, Tayi K. An introduction to glmnet. Available at: https://glmnet.stanford.edu/articles/glmnet.html. Accessed June 6, **2022**.
28. Friedman JH, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw **2010;** 33:1–22.
29. Friedman J, Hastie T, Tibshirani R, et al. Glmnet: lasso and elastic-net regularized generalized linear models. Available at: https://cran.r-project.org/web/packages/glmnet/glmnet.pdf. Accessed June 6, **2022**.
30. Pampalon R, Hamel D, Gamache P, Philibert MD, Raymond G, Simpson A. An area-based material and social deprivation index for public health in Québec and Canada. Can J Pub Health **2012**; 103(8 Suppl 2):S17–22.
31. The R Foundation, R Core Team. R: a language and environment for statistical computing. Available at: https://www.r-project.org. Accessed April 6, **2021**.
32. Kuhn M, Wing J, Weston S, et al. Caret: classification and regression training. Available at: https://cran.r-project.org/web/packages/caret/caret.pdf. Accessed June 6, **2022**.
33. Deer RR, Rock MA, Vasilevsky N, et al. Characterizing long COVID: deep phenotype of a complex condition. eBioMedicine **2021;** 74:103722.
34. Subramanian A, Nirantharakumar K, Hughes S, et al. Symptoms and risk factors for long COVID in non-hospitalized adults. Nat Med **2022;** 28: 1706–14.