Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

NLP-Assisted Case Identification and Interpretable Machine Learning for Long COVID Detection in Primary Care EMRs

Surani Matharaarachchi[†§,*], Alan Katz[‡], Mike Domaratzki[◇], Saman Muthukumarana[‡]
[†] New York Institute of Technology, Vancouver, BC, Canada
[‡] University of Manitoba, Winnipeg, MB, Canada
[◇] Western University, London, ON, Canada
[§] University of Victoria, Victoria, BC, Canada

Abstract

Identifying patients with Long COVID Syndrome (LCS) remains a challenge due to various symptoms, heterogeneous clinical presentation, and inconsistent documentation in electronic medical records. In this study, we develop a machine learning framework that uses natural language processing (NLP) to identify confirmed cases of LCS from physician encounter notes and to predict individuals at risk. Using data from the Manitoba COVID-19 Cohort linked to the Manitoba Primary Care Research Network (MaPCReN), we construct a set of characteristics that incorporate demographics, socioeconomic indicators, and pre and post-COVID symptom profiles. We frame Long COVID identification as an extreme class-imbalance NLP classification problem ($\sim 4\%$ confirmed cases in the development cohort) and address this challenge using imbalance-aware learning through random under-sampling and over-sampling strategies. Logistic regression with elastic net regularization combined with under-sampling achieves the best performance, with a sensitivity of 0.95, specificity of 0.81, and an AUC of 0.94, identifying 1,124 potential LCS cases among 4,556 COVID-19 positive individuals. These results demonstrate that combining unstructured clinical text with interpretable, imbalance-aware learning enables scalable Long COVID surveillance and risk identification in real-world EMR settings.

Keywords: Long COVID, Machine Learning, Elastic Net Classification, NLP, Class Imbalance, Electronic Medical Records

1. Introduction

Post-acute sequelae of SARS-CoV-2 infection (PASC), also referred to as Long COVID Syndrome (LCS), is a condition in which individuals experience prolonged symptoms weeks or months after recovering from the acute phase of COVID-19 [1]. Also known as post-COVID-19 condition (PCC), LCS affects both adults and children and can affect multiple organ systems, including respiratory, cardiovascular, neurological, and psychological functions [2, 3]. These long-lasting symptoms can severely reduce the quality of life of patients and limit their ability to perform daily activities.

Despite growing research efforts, LCS remains a clinically and biologically complex condition and its definition and mechanisms are still debated. Recent work shows that long COVID symptoms are highly heterogeneous across patients and do not fall into a few clear clusters, highlighting the continued need for personalized patient-centred care [4]. At the same time, many individuals with lingering symptoms may not be formally diagnosed, either due to lack of awareness, inconsistent clinical definitions, or limited access to care. A large cohort study by Huang et al. [3] found that approximately 20%–30% of individuals diagnosed with COVID-19 report symptoms lasting several months, underscoring the urgent need for robust, data-driven methods to identify and monitor LCS patients [5–9].

Previous studies have shown that certain patient characteristics and symptom profiles are associated with an increased risk of LCS, and have used data such as age, sex, socioeconomic status, and symptoms to train prediction models [5, 10, 11]. Sudre et al. used a random forest

* suranim@uvic.ca

model in mobile app data (AUC 0.76), while Pfaff et al. and Binka et al. developed XGBoost and elastic net models using structured EHRs and LCS clinic cohorts, achieving AUCs up to 0.93 and identifying key predictors such as fatigue, dyspnea, healthcare utilization, and age [5, 11, 12]. Other work has focused on symptom-based risk factors and phenotypes, using multivariable logistic regression and factor analysis to link LCS fatigue with autonomic dysfunction and cognitive symptoms and to identify symptom patterns [13, 14]. Text-mining approaches have also been applied to hospital-based narrative data to characterize Long COVID clinical conditions, demonstrating the potential of NLP for LCS phenotyping [15]. However, most existing models rely on predefined case labels from LCS clinics or hospitalized cohorts and do not explicitly address class imbalance. In our setting, no centralized long COVID clinic was available to provide confirmed diagnoses. To address this gap, we use NLP in primary care encounter notes in a population-based COVID-19 cohort to construct confirmed LCS labels and then train interpretable, imbalance-aware supervised models based on demographic, clinical, and symptom features. Our goal is to provide a robust and scalable approach for the early identification of at-risk individuals that can support clinical decision-making and public health planning in routine care settings.

Contributions. This work makes three contributions: (1) we propose an NLP-assisted case identification pipeline to construct confirmed Long COVID labels from primary care encounter notes in the absence of centralized Long COVID clinics; (2) we formulate Long COVID identification as an extreme class-imbalance learning problem and systematically compare interpretable classifiers with re-sampling strategies; and (3) we demonstrate population-level application by deploying the trained model to an unlabeled COVID-19-positive cohort to identify individuals at potential risk.

2. Materials and Methods

2.1. Data Collection

This study used data from the Manitoba COVID-19 Cohort, which includes all adult individuals tested for SARS-CoV-2 and reported COVID-19 cases in Manitoba between March 1, 2020, and December 31, 2021. A subset of COVID-19 positive individuals, including their demographics such as age, sex, and Socioeconomic Factor Index (SEFI), was obtained from the Manitoba Population Research Data Repository housed at the Manitoba Centre for Health Policy (MCHP) [16]. SEFI, derived from Census data, serves as a proxy for socioeconomic status (SES), incorporating variables such as average household income, percentage of single-parent households, unemployment rate (age 15+), and high school education rate [17].

These data were linked with electronic medical records (EMRs), including physicians' encounter notes, obtained from The Manitoba Primary Care Research Network (MaPCReN), to extract relevant LCS symptoms. MaPCReN is the Manitoba subnetwork of the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) [18]. The encounter notes were available from March 1, 2020, to September 28, 2022. However, some regional health authorities (RHAs) had incomplete participation. For example, Interlake-Eastern RHA joined MaPCReN in 2021, and Northern RHA has not contributed data since 2020. Furthermore, Winnipeg RHA (WRHA), which covers 57% of Manitoba's population, does not permit data sharing from its directly managed clinics. Table 1 describes the limitation of the representativeness of the sample. The percentage of the population for each RHA in 2021 is as reported in the Manitoba Health Population Report - June 1, 2021. These coverage limitations may introduce selection bias and reduce representativeness, so the results should be interpreted as applying to the MaPCReN-covered population rather than to all Manitobans.

Table 1. MaPCReN patients with encounter note records in April-June 2022

*RHA	Number of encounter notes	Number of patients with ≥ 1 note	With no encounter note records	Number of patients	Population % for RHA (2021)
IERHA	8,194	3,344	6,017	9,361	9.7%
PMRHA	9,913	2,763	9,122	11,885	12.4%
SRHA	57,891	22,529	56,561	79,090	15.5%
WRHA	72,881	28,954	143,690	172,644	56.9%
NRHA	-	-	-	-	5.5%
Total	148,879	57,590 (21.1%)	215,390 (78.9%)	272,980	100%

*RHA - Regional Health Authority, IERHA - Interlake-Eastern RHA, PMRHA - Prairie Mountain RHA, SRHA - Southern Health-Santé Sud RHA, WRHA - Winnipeg RHA, NRHA - Northern RHA

2.2. **Study Sample**

The eligibility criteria for this analysis were based on patients who had received a COVID-19 index date (the first date reported positive test result) from March 1, 2020, to December 31, 2021. After restricting to individuals where EMRs were available, the data set was narrowed to 4556 COVID-19 positive patients with written medical records. Evaluation of post-COVID symptoms began 90 days after the COVID-19 index date according to the WHO definition of LCS. In contrast, the observation period for assessing pre-COVID symptoms was expanded to encompass two scenarios: one considering symptoms within 90 days before the COVID-19 index date and the other examining symptoms within one year before the COVID-19 index date. This timeline allowed for the evaluation of symptoms before and after the onset of COVID-19, providing a comprehensive understanding of the symptom profile during the specified observation periods.

2.3. **Identifying the known LCS group**

The capacity of a computer to comprehend human language, whether spoken or written, is known as Natural Language Processing (NLP) [19, 20]. This manuscript uses a word extraction method to identify EMRs that specifically discuss LCS. As understanding of LCS evolves, it becomes increasingly important to extract and analyze relevant information from medical records efficiently. Manual review of all encounter notes can be a time-consuming and labour-intensive process, necessitating the need for automated techniques. The word extraction method combines NLP techniques with domain-specific knowledge to effectively identify encounter notes containing LCS information.

The word extraction method we employed involves a multi-step process. Initially, we compiled a comprehensive list of LCS-related keywords through an extensive literature review and expert input. These keywords encompass various synonyms that explain LCS, such as Long COVID, Post COVID, and Long Hauler. Subsequently, we developed a custom algorithm that uses NLP techniques, including tokenization, stemming, and regular expressions, to scan the encounter notes and detect instances of these LCS-related keywords.

In Fig. 1, the illustration depicts the refinement process of the established LCS group by using encounter notes. In step 3, we identified 121 patients with probable LCS-related encounter notes. In step 5, an expert review is introduced to further validate the NLP-filtered encounter notes. Of the 121 patients identified, 81 were identified as confirmed LCS patients, eight were determined to be unrelated to LCS, and the remaining 32 were categorized as having unknown conditions and subsequently added to the application dataset.

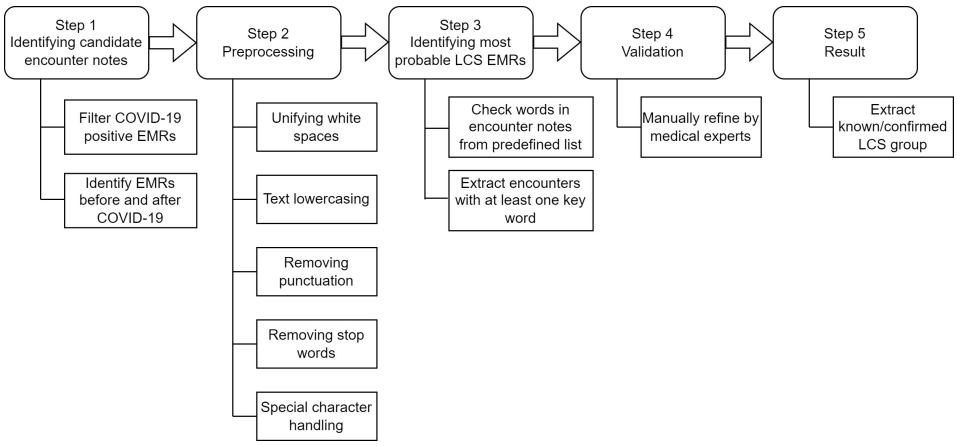

Figure 1. Identifying confirmed LCS group. LCS - Long COVID Syndrome, EMRs - Electronic Medical Records.

2.4. **Identifying the control group**

According to Amin-Chowdhury and Ladhani [9], the lack of appropriate controls is a significant limitation for designing machine learning tools, especially in assessing longer-term sequelae. Some studies have included people with symptomatic but negative COVID-19 test as control participants [5]. Some jurisdictions have LCS clinics that can help identify training data; for example, in a study with a cohort in British Columbia, the two classes provided to the algorithm were patients in the LCS clinic and patients who did not attend the LCS clinic [12]. Some others have compared patients with persistent symptoms with short-lasting illnesses.

In this study, we implemented an approach to identify a control group of patients who did not show LCS symptoms. Specifically, we defined the control group as comprising 1945 patients who remained within the dataset for at least 90 days without documented medical records beyond 90 days from the onset of COVID-19 (Fig. 2). This methodology is consistent with previous investigations that examined patients with LCS and control groups [21]. Our rationale for selecting patients without LCS symptoms was to establish a cohort of individuals with a comparable COVID-19 disease trajectory to those without long-term symptoms. This facilitated the isolation of LCS-specific symptoms and patient characteristics. Moreover, our selection criteria for patients who remained within the dataset for at least 90 days ensured adequate follow-up data to accurately classify patients as LCS or no LCS.

In general, our approach to identifying a control group aimed to minimize confounding factors, ensuring that any associations observed between patient attributes, symptoms, and LCS are specific to LCS and not merely reflective of general severity of COVID-19 disease. Additionally, we identified 2611 patients with medical records that extend beyond 90 days from the onset of COVID as potential patients with LCS. Among this group, our method effectively identified 222 encounter notes belonging to 121 unique patients mentioning LCS by leveraging the linguistic patterns and medical domain knowledge (Fig. 2).

2.5. **Symptom Extraction and Negation Identification**

To improve our identification of symptoms related to LCS in medical data, we harnessed the power of natural language processing (NLP) and specifically focused on detecting negation. We used Python-based tools to streamline these processes. We determined whether

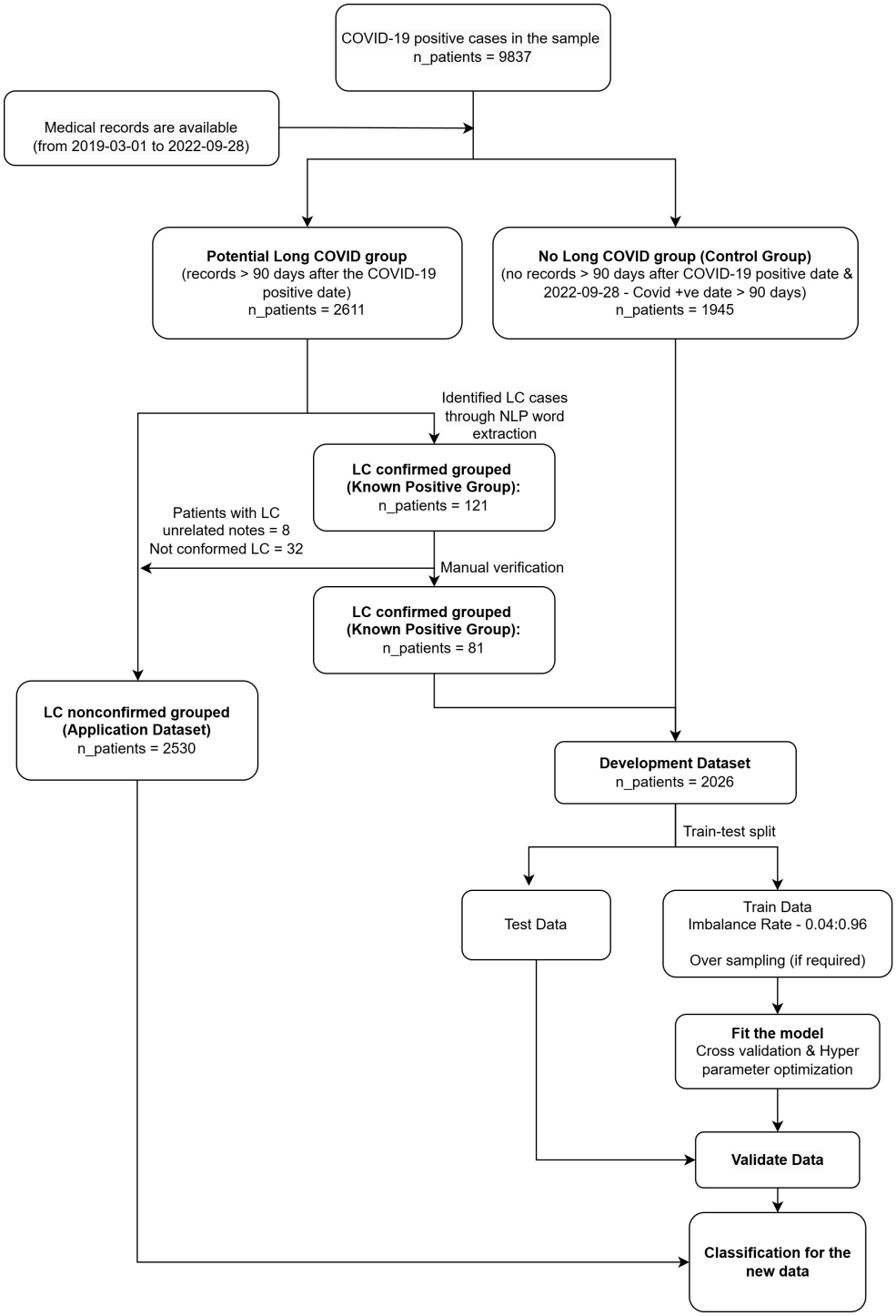

Figure 2. Process map

the words mentioned were affirmed or negated with *Negex* (Negation Extraction) [22] and excluded those with negated LCS symptoms. The *Negex* algorithm, a widely used algorithm for negation detection in clinical texts, identifies phrases such as “no", “not," or “but" and subsequently identifies any medical terms that follow as negated [22]. We also provided a separate list of terms that indicate negation (e.g., “abstain from," “other than," “excluding"). *Negex* examines the presence of these terms and carefully assesses the context to decide whether a given concept should be marked as negated. Using *Negex* allowed us to extract all negated medical terms from the electronic medical records (EMRs) of patients diagnosed with COVID-19. Concurrently, we extracted non-negated LCS-related symptoms by referring to a predetermined list of LCS symptoms as listed in [23].

However, for words such as taste, smell, appetite, and sleep, we are interested in the negation as loss of taste, smell, appetite, and lack of sleep were the symptoms of interest. In such cases, considering negations, we could more precisely identify the presence or absence of symptoms in each patient's EMR.

2.6. **Machine Learning Approach**

To predict the LCS status of patients given several patient attributes, we used supervised machine learning. In supervised machine learning, a Development Dataset is used to train and test the predictive model. The Development Dataset set consists of labeled examples in which the LCS status of patients is known. The patient characteristics, such as age, gender, socioeconomic information, and pre and post-COVID symptom profiles, are used as input features for the model. We split the Development Dataset into training and test sets, where the model learns patterns and relationships from the training data, allowing it to make predictions on new, unseen test data.

During training, the model adjusts its internal parameters to minimize the difference between the predicted LCS status and the actual labeled data. This process is often referred to as optimization or model fitting. Once trained, the model is applied to unseen test data to validate model performance and to new patient data (Application dataset) to predict their LCS status based on their attributes.

It is important to note that the performance of the predictive model depends on the quality and representativeness of the training data, as well as the choice of the machine learning algorithm and its associated parameters. This study used three binary classification methods: Logistic Regression, Logistic Regression with Elastic Net Regularization for Classification, and Random Forest Classification. These methods were implemented using *glmnet* and *caret* R packages. In addition, we used cross-validation and hyper-parameter optimization techniques to find the optimal set of parameters for each method.

Class imbalance is an important issue in predictive modeling, especially when one class is much smaller than the other [24]. In our study, LCS patients form a clear minority among all COVID-19 cases, so trained models naively tend to favour the no-LCS class. To address this, we used two simple re-sampling strategies: Random Over-Sampling, which duplicates LCS cases, and Random Under-Sampling, which removes no-LCS cases. These methods helped us move towards a 50%:50% class balance, allowing the models to learn minority patterns more effectively. While over-sampling can increase the risk of overfitting and under-sampling can discard information, in combination with regularized and tree-based classifiers they provided a good trade-off between sensitivity to LCS and overall model robustness.

Sensitivity, specificity, and the area under the ROC curve (AUC) were used to evaluate model performance.

Table 2. Identified LCS patient counts and percentages with model accuracy measures

Pre-COVID symptom scenario	Dataset	Re-sampling Technique	Classification Method	No LCS Counts (%)	LCS Counts (%)	Total LCS Counts (%) (Development + Application)	AUC	Sensitivity	Specificity
90 days	Development Dataset			1945 (96%)	81 (4%)				
	Application Dataset	Baseline (Without Re-sampling)	Logistic	1657 (65%)	873 (35%)	954 (20.9%)	0.87	0.85	0.82
			Elastic Net	1857 (73%)	673 (27%)	754 (16.5%)	0.93	0.85	0.91
			Random Forest	1656 (65%)	874 (35%)	955 (21%)	0.93	0.9	0.85
		Random Over-Sampling	Logistic	1912 (76%)	618 (24%)	699 (15.3%)	0.88	0.85	0.86
			Elastic Net	1689 (67%)	841 (33%)	922 (20.2%)	0.93	0.9	0.83
			Random Forest	1779 (70%)	751 (30%)	832 (18.3%)	0.9	0.85	0.84
		Random Under-Sampling	Logistic	1480 (58%)	1050 (42%)	1131 (24.8%)	0.66	0.7	0.71
			Elastic Net	**1487 (59%)**	**1043 (41%)**	**1124 (24.7%)**	**0.94**	**0.95**	**0.81**
			Random Forest	1659 (66%)	871 (34%)	952 (20.9%)	0.93	0.9	0.86
1 year	Development Dataset			1592 (95%)	81 (5%)				
	Application Dataset	Baseline (Without Re-sampling)	Logistic	1825 (72%)	705 (28%)	786 (18.7%)	0.69	0.69	0.88
			Elastic Net	1459 (58%)	1071 (42%)	1152 (27.4%)	0.86	0.85	0.84
			Random Forest	1225 (48%)	1305 (52%)	1386 (33%)	0.84	0.85	0.79
		Random Over-Sampling	Logistic	1626 (64%)	904 (36%)	985 (23.4%)	0.66	0.69	0.79
			Elastic Net	1753 (69%)	777 (31%)	858 (20.4%)	0.75	0.77	0.84
			Random Forest	1347 (53%)	1183 (47%)	1264 (30.1%)	0.87	0.85	0.79
		Random Under-Sampling	Logistic	1594 (63%)	936 (37%)	1017 (24.2%)	0.79	0.69	0.84
			Elastic Net	1621 (64%)	909 (36%)	990 (23.6%)	0.79	0.85	0.83
			Random Forest	1816 (72%)	714 (28%)	795 (18.9%)	0.89	0.85	0.9

3. **Results**

In our study, we evaluate three distinct machine learning algorithms alongside two alternative re-sampling techniques commonly used in the existing literature to predict individuals at risk of Long COVID Syndrome (LCS). Additionally, we consider two scenarios for including pre-COVID symptoms, one involving pre-COVID symptoms within 90 days of the COVID index date and another encompassing one year before the COVID index date. First, we divided our data set into training and test sets (80%:20%) to develop and evaluate our classification model for identifying LCS patients using patient attributes and symptoms. To perform the training-test split while preserving the class distribution in the response variable, we used the *createDataPartition* function from the *caret* package in R.

We then used 5-fold cross-validation to evaluate the model's performance further and mitigate over-fitting. The imbalance rate of the original data set was 4%:96% for the known LCS group and the no LCS control group (Fig. 2). We trained our model on the training set using classification algorithms and evaluated their performance on the test set. We used the *ROCR* package in R to establish a predictive probability threshold that maximized sensitivity and specificity. The best model was selected on the basis of specificity, sensitivity, and AUC.

After analyzing the performance of the nine selected models based on sensitivity and specificity, we identified the best model by prioritizing sensitivity as the primary criterion. Accurately detecting at-risk LCS patients is of utmost importance, and maximizing sensitivity ensures a higher probability of correctly identifying individuals susceptible to LCS. By selecting the model that optimizes sensitivity, we aim to minimize false negatives and enhance the model's ability to capture as many true positive cases as possible.

Table 2 illustrates that the random under-sampled logistic regression model with elastic net regularization demonstrated the highest sensitivity among the selected models, making it the preferred choice for identifying at-risk LCS patients. Table 2 displays relevant counts and percentages, indicating how each method classified patients in the Application Dataset as LCS cases. It shows the total counts and percentages of the COVID-19-positive people projected to develop LCS. The selected method predicted 1043 (41%) of the patients from the application dataset as LCS. Along with the 81 (4%) initially known LCS patients from the development dataset, this accounts for 24.7% (1124 patients out of 4556) of patients within the COVID-19-positive group projected to develop LCS. This best-fitted model chooses post-COVID symptoms such as Breathing/lung issues, fatigue, other pains, chest pain, brain fog,

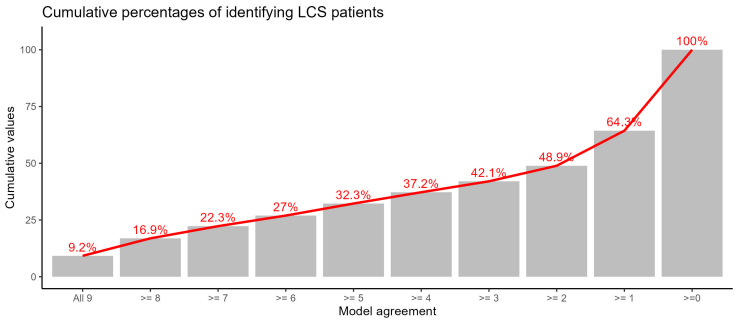

Figure 3. Agreement between nine models

Table 3. Demographic information of each patient group

Attribute	Development Dataset		Application Dataset		Prevalence Ratio		Overall	
	No LCS (Controls)	Known LCS	Predicted No LCS	Predicted LCS	No LCS	LCS	No LCS	LCS
	(N=1945)	(N=81)	(N=1487)	(N=1043)			(N=3432)	(N=1124)
Sex								
Female	1065 (54.8%)	45 (55.6%)	835 (56.2%)	626 (60.0%)	1.0	0.9	1900 (55.4%)	671 (59.7%)
Male	880 (45.2%)	36 (44.4%)	652 (43.8%)	417 (40.0%)	1.0	1.1	1532 (44.6%)	453 (40.3%)
SEFI								
Mean (SD)	-0.203 (0.680)	-0.191 (0.459)	-0.122 (0.724)	-0.101 (0.658)	1.7	1.9	-0.168 (0.701)	-0.107 (0.646)
Median [Min, Max]	-0.251 [-3.03, 3.16]	-0.178 [-1.98, 1.14]	-0.178 [-2.69, 3.88]	-0.178 [-3.40, 3.27]			-0.242 [-3.03, 3.88]	-0.178 [-3.40, 3.27]
Age Group								
<18	438 (22.5%)	s (s%)	236 (15.9%)	66 (6.3%)	1.4	-	674 (19.6%)	68 (6.0%)
18-29	367 (18.9%)	7 (8.6%)	301 (20.2%)	106 (10.2%)	0.9	0.8	668 (19.5%)	113 (10.1%)
30-39	323 (16.6%)	7 (8.6%)	225 (15.1%)	165 (15.8%)	1.1	0.5	548 (16.0%)	172 (15.3%)
40-49	259 (13.3%)	11 (13.6%)	198 (13.3%)	151 (14.5%)	1.0	0.9	457 (13.3%)	162 (14.4%)
50-59	253 (13.0%)	23 (28.4%)	215 (14.5%)	188 (18.0%)	0.9	1.6	468 (13.6%)	211 (18.8%)
60-69	167 (8.6%)	15 (18.5%)	160 (10.8%)	145 (13.9%)	0.8	1.3	327 (9.5%)	160 (14.2%)
70-79	74 (3.8%)	15 (18.5%)	79 (5.3%)	121 (11.6%)	0.7	1.6	153 (4.5%)	136 (12.1%)
80+	64 (3.3%)	s (s%)	73 (4.9%)	101 (9.7%)	0.7	-	137 (4.0%)	102 (9.1%)

[a] Prevalence ratio compares the prevalence among known vs predicting LCS patients (known/predicted),
[b] 's' indicates the values below six which were suppressed.

dizziness and cough and age group 70-79 as the most important eight features predicting the at-risk LCS group.

Following the manual refinement phase after the NLP approach, there were initial reservations about labelling 32 patients as LCS patients. Nevertheless, post-modeling, it became evident that the Elastic Net regression model, with under-sampling, accurately classified 27 of them, representing 84%, as LCS patients.

Fig. 3 shows a unanimous agreement among all nine models when using the 90-day pre-COVID window. There is notable consistency, with at least seven models jointly predicting 22.3 percent of patients within the COVID-19-positive group who are poised to transition into LCS.

Overall within the LCS group, 59.7% were female, compared with 55.4% in the No LCS group; the most common age category among LCS patients was 50–59 years (18.8%) (Table 3). These findings highlight potential associations between demographic factors and the appearance of LCS in the dataset.

Table 4 summarizes symptom prevalence (counts and percentages) in the post-COVID window by group. Table 5 contrasts pre- versus post-COVID prevalence and reports the absolute percentage-point change (post minus pre) over the 90-day window.

For most symptoms, prevalence was higher post-COVID than pre-COVID in both groups; however, the increase was consistently larger among LCS cases, particularly for respiratory symptoms, pain, fatigue, and cough (Table 5). These descriptive patterns align with the symptom features prioritized by the final elastic-net model.

Table 4. Post-COVID symptom classification of each patient group

Symptom	Development Dataset		Application Dataset		Prevalence Ratio	
	No LCS Controls	Known LCS	Predicted No LCS	Predicted LCS	No LCS	LCS
	(N=1945)	(N=81)	(N=1487)	(N=1043)		
Ache	110 (5.7%)	16 (19.8%)	99 (6.7%)	188 (18.0%)	0.9	1.1
Allergy	132 (6.8%)	16 (19.8%)	142 (9.5%)	167 (16.0%)	0.7	1.2
Appetite	32 (1.6%)	s (s%)	32 (2.2%)	48 (4.6%)	0.7	-
Arm pain	166 (8.5%)	18 (22.2%)	207 (13.9%)	286 (27.4%)	0.6	0.8
Back pain	137 (7.0%)	12 (14.8%)	114 (7.7%)	199 (19.1%)	0.9	0.8
Body pain	13 (0.7%)	0 (0%)	7 (0.5%)	27 (2.6%)	1.4	0.0
Breathing/lung issues	325 (16.7%)	63 (77.8%)	0 (0%)	845 (81.0%)	-	1.0
Cough	346 (17.8%)	47 (58.0%)	221 (14.9%)	573 (54.9%)	1.2	1.1
Dysphagia	16 (0.8%)	s (s%)	16 (1.1%)	33 (3.2%)	0.7	-
Hand pain	s (s%)	0 (0%)	s (s%)	17 (1.6%)	-	0.0
Headache	185 (9.5%)	26 (32.1%)	168 (11.3%)	309 (29.6%)	0.8	1.1
Heart	163 (24%)	31 (38.3%)	146 (9.8%)	319 (30.6%)	0.9	1.3
Hemoptysis	s (s%)	s (s%)	s (s%)	25 (2.4%)	-	-
Hot flushes	s (s%)	0 (0%)	9 (0.6%)	8 (0.8%)	-	0.0
Neck	145 (7.5%)	25 (30.9%)	154 (10.4%)	227 (21.8%)	0.7	1.4
Pain	708 (36.4%)	64 (79.0%)	723 (48.6%)	811 (77.8%)	0.7	1.0
Phlegm	27 (1.4%)	11 (13.6%)	15 (1.0%)	68 (6.5%)	1.4	2.1
Shoulder pain	32 (1.6%)	s (s%)	36 (2.4%)	61 (5.8%)	0.7	-
Trauma	47 (2.4%)	8 (9.9%)	87 (5.9%)	116 (11.1%)	0.4	0.9
Weight loss	54 (2.8%)	8 (9.9%)	68 (4.6%)	116 (11.1%)	0.6	0.9
Chest pain	240 (12.3%)	42 (51.9%)	134 (9.0%)	520 (49.9%)	1.4	1.0
Fatigue	180 (9.3%)	51 (63.0%)	54 (3.6%)	473 (45.4%)	2.6	1.4
Brain fog	52 (2.7%)	19 (23.5%)	65 (4.4%)	123 (11.8%)	0.6	2.0
Hoarse voice	s (s%)	s (s%)	s (s%)	16 (1.5%)	-	-
Ear pain	63 (3.2%)	9 (11.1%)	65 (4.4%)	84 (8.1%)	0.7	1.4
Joint pain	79 (4.1%)	10 (12.3%)	87 (5.9%)	148 (14.2%)	0.7	0.9
Muscle pain	84 (4.3%)	12 (14.8%)	82 (5.5%)	174 (16.7%)	0.8	0.9
Mouth ulcer	s (s%)	s (s%)	9 (0.6%)	14 (1.3%)	-	-
Numbness	70 (3.6%)	10 (12.3%)	84 (5.6%)	135 (12.9%)	0.6	1.0
Leg pain	195 (10.0%)	26 (32.1%)	164 (11.0%)	283 (27.1%)	0.9	1.2
Loss smell	22 (1.1%)	s (s%)	20 (1.3%)	33 (3.2%)	0.8	-
Loss taste	16 (0.8%)	s (s%)	19 (1.3%)	28 (2.7%)	0.6	-
Dizziness	128 (6.6%)	26 (32.1%)	120 (8.1%)	296 (28.4%)	0.8	1.1

[a] Prevalence ratio compares the prevalence among known vs predicting LCS patients (known/predicted),
[b] 's' indicates the values below six which were suppressed.

4. **Discussion**

The primary objective of this study was to develop a computational phenotyping model capable of accurately identifying Long COVID Syndrome (LCS) cases within a large dataset of reported COVID-19 cases in Manitoba, Canada. To achieve this, we evaluated three supervised machine learning methods: logistic regression, logistic regression with elastic net regularization, and random forest classification. The training dataset was constructed using individuals identified as confirmed LCS patients using a natural language processing (NLP) keyword extraction approach, with features derived from demographics, socioeconomic indicators (SEFI), and pre- and post-COVID symptom profiles.

Our goal was to create a robust and interpretable model using routinely collected healthcare data under class imbalance. To address this challenge, we applied random over-sampling and random under-sampling strategies. Among the evaluated models, logistic regression with elastic net regularization achieved the best performance when combined with under-sampling, identifying 1,124 LCS cases and representing 24.7% of the 4,556 COVID-19 positive individuals.

From a machine learning perspective, these results suggest that under class imbalance, regularized linear models can generalize effectively when the predictive signal is sparse

Table 5. Overall classification of symptoms with percentage difference

Symptom	No LCS			LCS		
	Pre-COVID	Post-COVID	Percentage Difference	Pre-COVID	Post-COVID	Percentage Difference
Ache	29 (0.8%)	209 (6.1%)	5.3%	16 (1.4%)	204 (18.1%)	16.7%
Allergy	38 (1.1%)	274 (8.0%)	6.9%	22 (2.0%)	183 (16.3%)	14.3%
Appetite	10 (0.3%)	64 (1.9%)	1.6%	s (s%)	52 (4.6%)	-
Arm pain	44 (1.3%)	373 (10.9%)	9.6%	45 (4.0%)	304 (27.0%)	23%
Back pain	27 (0.8%)	251 (7.3%)	6.5%	43 (3.8%)	211 (18.8%)	15%
Body pain	s (s%)	20 (0.6%)	-	s (s%)	27 (2.4%)	-
Breathing/lung issues	91 (2.7%)	325 (9.5%)	6.8%	106 (9.4%)	908 (80.8%)	71.4%
Cough	103 (3.0%)	567 (16.5%)	13.5%	97 (8.6%)	620 (55.2%)	46.6%
Dysphagia	6 (s%)	32 (0.9%)	-	8 (0.7%)	37 (3.3%)	2.6%
Hand pain	0 (0%)	7 (0.2%)	0.2%	s (s%)	17 (1.5%)	-
Headache	61 (1.8%)	353 (10.3%)	8.5%	72 (6.4%)	335 (29.8%)	23.4%
Heart	47 (1.4%)	309 (9.0%)	7.6%	39 (3.5%)	350 (31.1%)	27.6%
Hemoptysis	s (s%)	9 (0.3%)	-	s (s%)	26 (2.3%)	-
Hot flushes	s (s%)	14 (0.4%)	-	0 (0%)	8 (0.7%)	0.7%
Neck	33 (1.0%)	299 (8.7%)	7.7%	37 (3.3%)	252 (22.4%)	19.1%
Pain	264 (7.7%)	1431 (41.7%)	34%	222 (19.8%)	875 (77.8%)	58%
Phlegm	9 (0.3%)	42 (1.2%)	0.9%	13 (1.2%)	79 (7.0%)	5.8%
Shoulder pain	s (s%)	68 (2.0%)	-	11 (1.0%)	67 (6.0%)	5%
Trauma	20 (0.6%)	134 (3.9%)	3.3%	18 (1.6%)	124 (11.0%)	9.4%
Weight loss	16 (0.5%)	122 (3.6%)	3.1%	14 (1.2%)	124 (11.0%)	9.8%
Chest pain	74 (2.2%)	374 (10.9%)	8.7%	69 (6.1%)	562 (50.0%)	43.9%
Fatigue	47 (1.4%)	234 (6.8%)	5.4%	61 (5.4%)	524 (46.6%)	41.2%
Brain fog	13 (0.4%)	117 (3.4%)	3%	18 (1.6%)	142 (12.6%)	11%
Hoarse voice	s (s%)	11 (0.3%)	-	s (s%)	17 (1.5%)	-
Ear pain	15 (0.4%)	128 (3.7%)	3.3%	9 (0.8%)	93 (8.3%)	7.5%
Joint pain	22 (0.6%)	166 (4.8%)	4.2%	17 (1.5%)	158 (14.1%)	12.6%
Muscle pain	21 (0.6%)	166 (4.8%)	4.2%	32 (2.8%)	186 (16.5%)	13.7%
Mouth ulcer	s (s%)	10 (0.3%)	-	s (s%)	15 (1.3%)	-
Numbness	28 (0.8%)	154 (4.5%)	3.7%	21 (1.9%)	145 (12.9%)	11%
Leg pain	52 (1.5%)	359 (10.5%)	9%	49 (4.4%)	309 (27.5%)	23.1%
Loss smell	8 (0.2%)	42 (1.2%)	1%	11 (1.0%)	35 (3.1%)	2.1%
Loss taste	s (s%)	35 (1.0%)	-	9 (0.8%)	31 (2.8%)	2%
Dizziness	34 (1.0%)	248 (7.2%)	6.2%	48 (4.3%)	322 (28.6%)	24.3%

[a] Prevalence ratio compares the prevalence among known vs predicting LCS patients (known/predicted),

[b] 's' indicates the values below six which were suppressed.

and correlated. Elastic net regularization provided stability under multicollinearity while performing feature selection, enabling transparent identification of a concise set of predictive symptom patterns. Random under-sampling further improved minority-class detection by reducing majority-class dominance during training, facilitating more balanced decision boundaries.

These properties make the elastic net particularly well suited for LCS identification in this setting. By combining L1 and L2 penalties, elastic net balances interpretability, robustness, and generalization, while mitigating overfitting in high-dimensional and correlated feature spaces. This aligns with prior findings that highlight the effectiveness of elastic net models in clinical risk prediction under class imbalance [5].

We acknowledge several limitations. EMR data was obtained from physicians who consented to data sharing, which can introduce selection bias and limit population coverage. Furthermore, the absence of centralized Long COVID clinics in Manitoba required reliance on NLP-assisted case identification, which can introduce label noise. A fully population-based study incorporating all EMRs, along with external validation across jurisdictions, would strengthen generalizability. Future work will explore richer text representations and calibrated decision thresholds to further support deployment.

As additional data sources become available, the proposed framework can be retrained and extended to improve LCS detection in diverse clinical and population settings.

5. Conclusion

This study presents a scalable and practical approach to LCS identification by combining natural language processing to detect confirmed LCS cases with supervised machine learning models trained in comprehensive patient attributes. The logistic regression model with elastic net regularization, in combination with random under-sampling, achieved high predictive performance, with a sensitivity of 0.95, specificity of 0.81, and an AUC of 0.94. These results demonstrate the feasibility of using routinely collected clinical data and interpretable machine learning to support early identification and monitoring of LCS patients.

The identified LCS cohort serves as a valuable foundation for future clinical and epidemiological studies and supports data-driven decision-making in healthcare. As additional data become available, this framework can be adapted and expanded to improve LCS detection and management across broader populations.

Acknowledgements

Muthukumarana has been partially supported by research grants from the Natural Sciences and Engineering Research Council of Canada (NSERC), and Katz has been supported by Canadian Institute of Health Research (CIHR). The authors acknowledge the Manitoba Centre for Health Policy (MCHP) for providing the data.

References

[1] A. Nalbandian et al. "Post-acute COVID-19 syndrome". eng. In: *Nature medicine* 27.4 (2021), pp. 601–615. ISSN: 1078-8956. DOI: https://doi.org/10.1038/s41591-021-01283-z.

[2] A. Carfì, R. Bernabei, and F. Landi. "Persistent Symptoms in Patients After Acute COVID-19". eng. In: *JAMA : the journal of the American Medical Association* 324.6 (2020), pp. 603–605. ISSN: 0098-7484. DOI: https://doi.org/10.1001/jama.2020.12603.

[3] C. Huang et al. "6-month consequences of COVID-19 in patients discharged from hospital: a cohort study". eng. In: *The Lancet (British edition)* 397.10270 (2021), pp. 220–232. ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(20)32656-8.

[4] E. Fresi, E. Pagani, F. Pezzetti, C. Montomoli, C. Monti, M. Betti, A. De Silvestri, O. Sagliocco, V. Zuccaro, R. Bruno, and C. Klersy. "Long COVID's Hidden Complexity: Machine Learning Reveals Why Personalized Care Remains Essential". In: *Journal of Clinical Medicine* 14.11 (2025). ISSN: 2077-0383. DOI: 10.3390/jcm14113670. URL: https://www.mdpi.com/2077-0383/14/11/3670.

[5] M. Binka, B. Klaver, G. Cua, A. W. Wong, C. Fibke, H. A. Velásquez García, P. Adu, A. Levin, S. Mishra, B. Sander, H. Sbihi, and N. Z. Janjua. "An Elastic Net Regression Model for Identifying Long COVID Patients Using Health Administrative Data: A Population-Based Study". eng. In: *Open forum infectious diseases* 9.12 (2022), ofac640–ofac640. ISSN: 2328-8957. DOI: https://doi.org/10.1093/ofid/ofac640.

[6] L. Bull-Otterson, S. Baca, S. Saydah, T. K. Boehmer, S. Adjei, S. Gray, and A. M. Harris. *Post–COVID Conditions Among Adult COVID-19 Survivors Aged 18–64 and ≥ 65 Years — United States, March 2020–November 2021*. eng. Atlanta, 2022.

[7] H. E. Davis, G. S. Assaf, L. McCorkell, H. Wei, R. J. Low, Y. Re'em, S. Redfield, J. P. Austin, and A. Akrami. "Characterizing Long COVID in an international cohort: 7 months of symptoms and their impact". eng. In: *EClinicalMedicine* 38 (2021), pp. 101019–101019. ISSN: 2589-5370. DOI: https://doi.org/10.1016/j.eclinm.2021.101019.

[8] H. Crook, S. Raza, J. Nowell, M. Young, and P. Edison. "Long COVID—mechanisms, risk factors, and management". eng. In: *BMJ (Online)* 374 (2021), n1648–n1648. ISSN: 1756-1833. DOI: https://doi.org/10.1136/bmj.n1648.

[9] Z. Amin-Chowdhury and S. N. Ladhani. "Causation or confounding: why controls are critical for characterizing Long COVID". eng. In: *Nature medicine* 27.7 (2021), pp. 1129–1130. ISSN: 1078-8956. DOI: https://doi.org/10.1038/s41591-021-01402-w.

[10] B. Blomberg, K. G.-I. Mohn, K. A. Brokstad, F. Zhou, D. W. Linchausen, B.-A. Hansen, and S. Lartey. "Long COVID in a prospective cohort of home-isolated patients". eng. In: *Nature medicine* 27.9 (2021), pp. 1607–1613. ISSN: 1078-8956. DOI: https://doi.org/10.1038/s41591-021-01433-3.

[11] C. H. Sudre et al. "Attributes and predictors of Long COVID". eng. In: *Nature Medicine* 27.4 (2021), pp. 626–631. ISSN: 1078-8956. DOI: https://doi.org/10.1038/s41591-021-01292-y.

[12] E. R. Pfaff et al. "Identifying who has Long COVID in the USA: a machine learning approach using N3C data". eng. In: *The Lancet (British edition)* 4.7 (2022), e532–e541. ISSN: 2589-7500. DOI: https://doi.org/10.1016/S2589-7500(22)00048-6.

[13] I. Margalit et al. "Risk Factors and Multidimensional Assessment of Long Coronavirus Disease Fatigue: A Nested Case-Control Study". eng. In: *Clinical infectious diseases* 75.10 (2022), pp. 1688–1697. ISSN: 1058-4838. DOI: https://doi.org/10.1093/cid/ciac283.

[14] D. Yelin et al. "Patterns of Long COVID Symptoms: A Multi-Center Cross Sectional Study". eng. In: *Journal of clinical medicine* 11.4 (2022), pp. 898–. ISSN: 2077-0383. DOI: https://doi.org/10.3390/jcm11040898.

[15] P. T. Veras Florentino et al. "Text mining method to unravel Long COVID's clinical condition in hospitalized patients". In: *Cell Death & Disease* 15.9 (2024), p. 671. DOI: 10.1038/s41419-024-07043-4.

[16] *A systematic investigation of Manitoba's Provincial Laboratory data*. eng. Winnipeg: Manitoba Centre for Health Policy, Dept. of Community Health Sciences, University of Manitoba, 2012. ISBN: 9781896489681.

[17] D. Chateau, C. Metge, H. Prior, and R.-A. Soodeen. "Learning From the Census: The Socioeconomic Factor Index (SEFI) and Health Outcomes in Manitoba". eng ; fre. In: *Canadian journal of public health* 103.8 Suppl 2 (2012), S23–S27. ISSN: 0008-4263. DOI: https://doi.org/10.1007/BF03403825.

[18] R. V. Birtwhistle. "Canadian Primary Care Sentinel Surveillance Network: a developing resource for family medicine and public health". eng. In: *Canadian family physician* 57.10 (2011), pp. 1219–1220. ISSN: 0008-350X. DOI: https://doi.org/10.1007/BF03403825.

[19] L. A. Celi, M. S. Majumder, P. Ordóñez, J. S. Osorio, K. E. Paik, and M. Somai. "Introduction to Clinical Natural Language Processing with Python". eng. In: *Leveraging Data Science for Global Health*. Switzerland: Springer International Publishing AG, 2020. ISBN: 3030479935.

[20] D. Keselj Vlado (Review of: Jurafsky and J. H. Martin. "Speech and Language Processing (second edition)". eng. In: *Computational Linguistics* 35.3 (2009), pp. 463–466. ISSN: 0891-2017. DOI: https://doi.org/10.1162/coli.B09-001.

[21] H. Moldofsky and J. Patcai. "Chronic widespread musculoskeletal pain, fatigue, depression and disordered sleep in chronic post-SARS syndrome; a case-controlled study". eng. In: *BMC neurology* 11.1 (2011), pp. 37–37. ISSN: 1471-2377. DOI: https://doi.org/10.1186/1471-2377-11-37.

[22] W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. "A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries". eng. In: *Journal of biomedical informatics* 34.5 (2001), pp. 301–310. ISSN: 1532-0464. DOI: https://doi.org/10.1006/jbin.2001.1029.

[23] S. Matharaarachchi, M. Domaratzki, A. Katz, and S. Muthukumarana. "Discovering Long COVID Symptom Patterns: Association Rule Mining and Sentiment Analysis in Social Media Tweets". eng. In: *JMIR formative research* 6.9 (2022), e37984–e37984. ISSN: 2561-326X. DOI: https://doi.org/10.2196/37984.

[24] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana. "Assessing feature selection method performance with class imbalance data". In: *Machine Learning with Applications* 6 (2021), p. 100170. ISSN: 2666-8270. DOI: https://doi.org/10.1016/j.mlwa.2021.100170. URL: https://www.sciencedirect.com/science/article/pii/S2666827021000852.