Uncovering Symptoms and Predicting Long COVID Using Social Media Tweets and Clinical Notes Data

A machine Learning Approach

Surani Matharaarachchi

Joint work with:
Dr. Saman Muthukumarana (University of Manitoba),
Dr. Alan Katz (University of Manitoba) &
Dr. Mike Domaratzki (Western University)

December, 29 2024

University of Manitoba

Introduction

Long COVID Syndrome (LCS)

- A condition in which individuals experience symptoms for weeks or months after recovering from COVID-19 [6].
- The need for consistent identification and treatment of Long COVID patients
  - 20-30% of COVID-19 survivors experience prolonged symptoms [2, 3].
  - The condition can affect multiple organ systems.
  - Many are unaware of their condition.

Objectives

1. To understand the patterns and behaviour of LCS symptoms reported by patients on the Twitter social media platform.
2. To establish a robust and precise model for identifying individuals experiencing LCS in Manitoba.

Symptom Pattern Recongnition

Discovering Long COVID Symptom Patterns: Association Rule Mining and Sentiment Analysis in Social Media Tweets [5]

Data Collection I

- Long COVID–related Twitter data were collected from May 1, 2020, to December 31, 2021.
- Data set of about 1M tweets.
- Used the Snscrape module in Python 3.8 [1] to scrape the tweet text online from tweets that match the keyword “LongCovid.”
- We reduced the data set to 127,848 tweets by limiting the population to those who suffered from COVID-19.

Pre-Processed Data

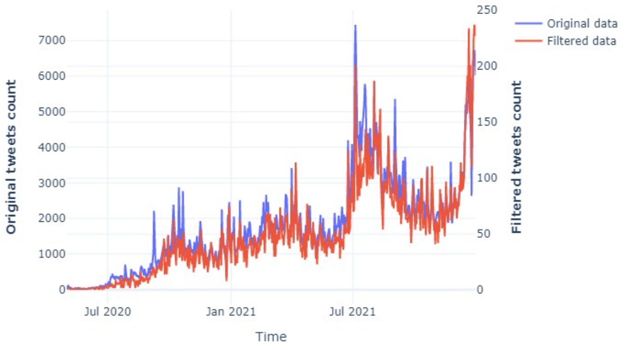


Figure: Time series plot for originally obtained data and the data considered for the study.

Natural Language Processing Techniques

- Tokenization
- Stopword Removal
- Stemming
- Sentiment Analysis
- Word Collocations
    - Pointwise Mutual Information (PMI)
    - t-test with a frequency filter
    - Chi-square test

Symptom Identification

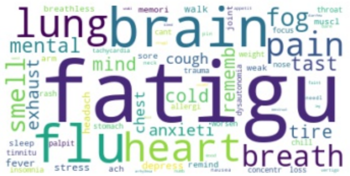


Figure: Word cloud of the list of 73 words after stemming.

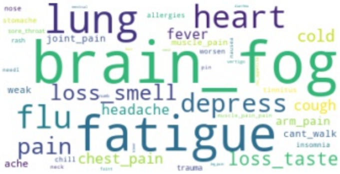
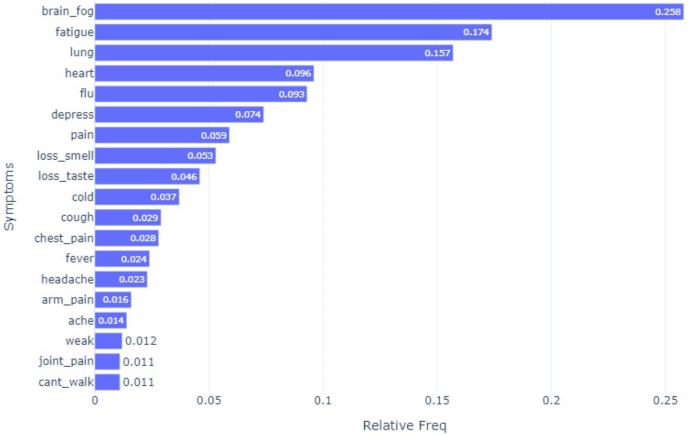


Figure: Word cloud of the list of 44 identified symptoms.

Relative Frequency of Symptoms in Long COVID Patients



Association Rule Mining Techniques

- Used association rule mining techniques to identify frequent symptoms and establish relationships between symptoms among patients with Long COVID in Twitter social media discussions.

- The highest confidence level–based detection was used to determine the most significant rules with 10% minimum confidence and 0.01% minimum support with a positive lift.

Association Rule Mining Techniques

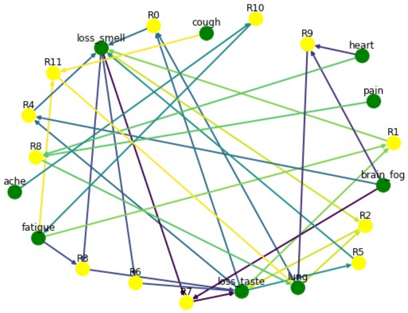


Figure: Association rules visualization. R: rule.

Predictive Models for LCS

Long COVID Prediction in Manitoba Using Clinical Notes Data [4]

Importance

Develop a computational predictive model to identify LCS cases precisely.

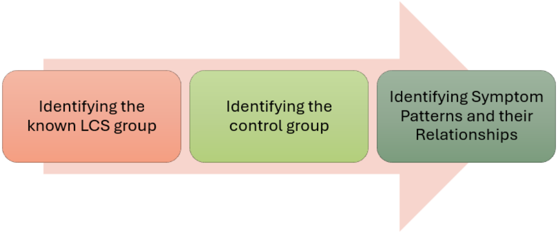
- Leveraging machine learning techniques offer a promising approach to accurately identifying and managing LCS cases.
- Potential to revolutionize the identification process of LCS, making it a significant contribution to the medical field.
- Improving patient care and management strategies.

Data Collection II

- Source of Data
  - Manitoba Population Research Data Repository housed at the Manitoba Centre for Health Policy (MCHP)
  - Electronic Medical Records (EMR) of COVID-19 patients pre- and post-COVID.
  - Demographic information such as age, sex, and socioeconomic factor index.
- Quantity and Quality
  - A sample of the COVID-19 test-positive cohort was accessed.
  - Patients who had received a COVID-19 index date from March 1, 2020, to December 31, 2021.
  - The data set was narrowed to 4556 COVID-19-positive patients with written medical records.
- Ethical Considerations
  - Data collection adhered to ethical guidelines, with measures in place to ensure patient privacy and confidentiality.

Challenges in Predicting LCS Patients at Risk

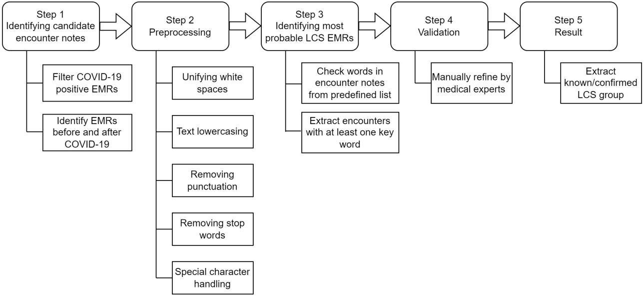
- The absence of a definitive diagnostic test for Long COVID Syndrome.
  - Identifying the known LCS Group for classification.
  - Defining the Control Group.



- Class imbalance issue

Identifying the known LCS Group

- Use Natural Language Processing (NLP) methodologies.
- Conducted word extraction processes.
- Out of 121 patients identified, 81 were confirmed LCS patients.



Defining the Control Group

- Who remained within the dataset for at least 90 days with no documented medical records beyond 90 days from COVID-19 onset.
- Identified 1945 patients.
- Class imbalance Ratio: 0.96:0.04

Symptom Extraction and Negation Identification

- Assessing post-COVID symptoms 90 days after the COVID-19 index date.
- Pre-COVID symptoms
  1. symptoms within 90 days before the COVID-19 index date.
  2. symptoms within one year before the COVID-19 index date.
- Extracted non-negated LCS-related symptoms by referring to a predetermined list [5].
  - Using ‘Negex’ allowed us to filter out all negated medical terms from the EMRs of patients.

Machine Learning Approach

- Supervised machine learning
- Train-test split
- Resampling Techniques
  - Random Over-sampling
  - Random Under-sampling
- Binary classification methods
  - Logistic Regression
  - Logistic Regression with Elastic Net Regularization for Classification
  - Random Forest Classification
- Cross-validation and hyperparameter optimization techniques

Model Results

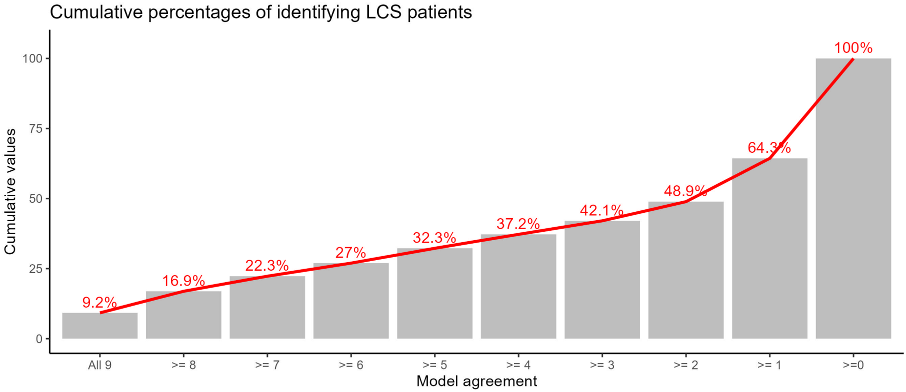
Table: Identified LCS patient counts and percentages with model accuracy measures

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

Model Results

- Pre-COVID Symptom Scenario: within 90 days prior to the COVID index date
- Logistic Regression with Elastic Net Regularization
- Random Under-Sampling
- AUC - 0.94, Sensitivity - 0.95, Specificity - 0.81
- Identified LCS group in Risk: 1124 (24.7%) LCS patients from the set of 4556 COVID-19 cases
- Most Important Features: Breathing/lung issues, Fatigue, Chest pain, Brain fog, Dizziness, Cough, Age group 70-79

Agreement between Nine Models



Discussion

- Using natural language processing to identify LCS symptoms patterns and initial confirmed LCS patients.
- Applying machine learning models addresses a significant challenge within the healthcare sector.
- The outcomes of this approach underscore its potential to accurately identify individuals prone to LCS, with higher accuracy.
- The LCS patient cohort created using this method is a valuable resource for conducting robust assessments of LCS clinical progression.

References

[1] Argamon, Shlomo (Review of: Pang, B. and L. Lee (2009). Opinion mining and sentiment analysis. *Computational Linguistics 35*(2), 311–312.

[2] Carfì, A., R. Bernabei, and F. Landi (2020). Persistent symptoms in patients after acute covid-19. *JAMA : the journal of the American Medical Association 324*(6), 603–605.

[3] Huang, C., L. Huang, Y. Wang, X. Li, L. Ren, X. Gu, L. Kang, L. Guo, M. Liu, X. Zhou, J. Luo, Z. Huang, S. Tu, Y. Zhao, L. Chen, D. Xu, Y. Li, C. Li, L. Peng, Y. Li, W. Xie, D. Cui, L. Shang, G. Fan, J. Xu, G. Wang, Y. Wang, J. Zhong, C. Wang, J. Wang, D. Zhang, and B. Cao (2021). 6-month consequences of covid-19 in patients discharged from hospital: a cohort study. *The Lancet (British edition) 397*(10270), 220–232.

[4] Matharaarachchi, S., M. Domaratzki, A. Katz, and S. Muthukumarana (2022). Discovering long covid symptom patterns: Association rule mining and sentiment analysis in social media tweets. *JMIR formative research 6*(9), e37984–e37984.

[5] Matharaarachchi, S., M. Domaratzki, A. Katz, and S. Muthukumarana (2024). Long covid prediction in manitoba using clinical notes data: A machine learning approach. *Intelligence-Based Medicine (In Review)*.

[6] Nalbandian, A., K. Sehgal, A. Gupta, M. V. Madhavan, C. McGroder, J. S. Stevens, J. R. Cook, A. S. Nordvig, D. Shalev, T. S. Sehrawat, N. Ahluwalia, B. Bikdeli, D. Dietz, C. Der-Nigoghossian, N. Liyanage-Don, G. F. Rosner, E. J. Bernstein, S. Mohan, A. A. Beckley, D. S. Seres, T. K. Choueiri, N. Uriel, J. C. Ausiello, D. Accili, D. E. Freedberg, M. Baldwin, A. Schwartz, D. Brodie, C. K. Garcia, M. S. V. Elkind, J. M. Connors, J. P. Bilezikian, D. W. Landry, and E. Y. Wan (2021). Post-acute covid-19 syndrome. *Nature medicine 27*(4), 601–615.

Thank You!

Contact: matharas@myumanitoba.ca