Machine Learning-Based Identification of Long COVID Syndrome

Leveraging Encounter Notes Symptoms

Surani Matharaarachchi

PhD Candidate, Department of Statistics, University of Manitoba

Data Scientist, Government of Manitoba

matharas@myumanitoba.ca

Joint work with Dr. Saman Muthukumarana, Dr. Mike Domaratzki, Dr. Alan Katz

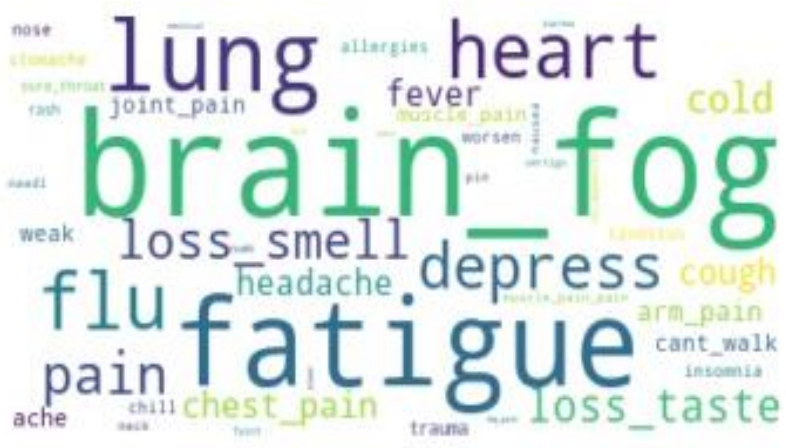
University of Manitoba

4th International Conference on Future of Preventive Medicine & Public Health
March 18-19, 2024
Virtual Event

Introduction

- Long COVID Syndrome (LCS)
  - A condition in which individuals experience symptoms for weeks or months after recovering from COVID-19 [1].
  - Post-COVID symptoms 90 days after the COVID-19 index date.

- The need for consistent identification and treatment of Long COVID patients
  - 20-30% of COVID-19 survivors experience prolonged symptoms.
  - The condition can affect multiple organ systems [2,3].
  - Many are unaware of their condition.



[4]

Objective

- To develop a computational predictive model to identify LCS cases precisely.

- Importance:
  - Leveraging machine learning techniques offers 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

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

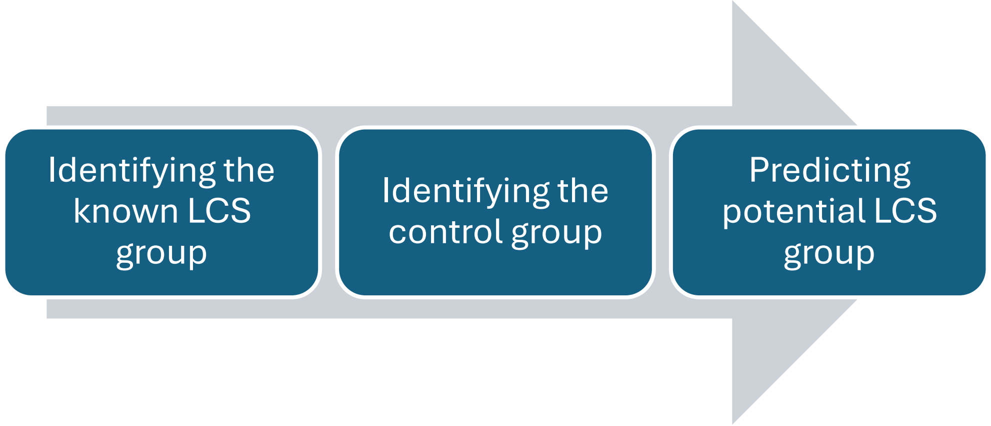
- A sample of the COVID-19 test-positive cohort was accessed.
- Patients who had received a COVID-19 index date from March 1, 2020, to December 31, 2021.
- The data set was narrowed to 4556 COVID-19-positive patients with written medical records.

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

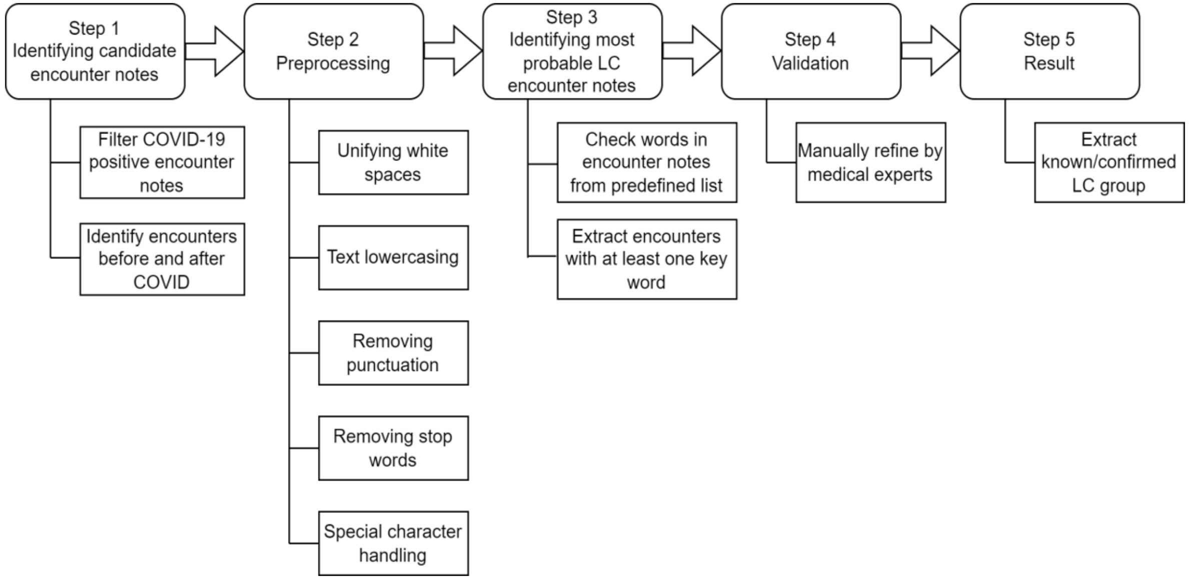
- The absence of a definitive diagnostic test for Long COVID Syndrome
  - Identifying the known LCS Group for classification
  - Defining the Control Group



- Class Imbalance Issue

Identifying the known LCS Group & Control Group

- 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 Issue

- One or more classes are underrepresented.
  - Class imbalance Ratio: 0.96:0.04

- Used resampling techniques
  - Random Over-Sampling
  - Random Under-Sampling

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 [4].
  - 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
- Binary classification methods
  - Logistic Regression
  - Logistic Regression with Elastic Net Regularization for Classification
  - Random Forest Classification
- Cross-validation and hyperparameter optimization techniques

Logistic Regression with Elastic Net Regularization

- Based on a linear combination of L1 and L2 regularization penalties, which are applied to the coefficients of the linear classification model.
- Elastic net regularization seeks to find coefficients that can minimize,

$$\min_{(b,w)\in\mathbb{R}^{\{m+1\}}}\left(-\frac{l(b,w)}{n}+\lambda P_{\alpha}(w)\right),\ \text{where, } P_{\alpha}(w)=(1-\alpha)\frac{1}{2}\left|\left|w\right|\right|_2^2+\alpha\left|\left|w\right|\right|_1$$

Also,

$$\left|\left|w\right|\right|_1=\sum_{j=1}^{p}|w_j| \quad \text{and} \quad \left|\left|w\right|\right|_2=\left(\sum_{j=1}^{p}w_j^2\right)^{\frac{1}{2}},\quad \alpha\geq 0,\ \ \lambda\in[0,1]$$

Accuracy Measures

- Assessed the overall performance of LCS prediction models by striking a balance between sensitivity and specificity.
- Sensitivity:
  - model's ability to correctly detect individuals with the disease.

$$Sensitivity = \frac{TP}{(TP + FN)}$$

- Specificity:
  - model's ability to accurately classify individuals who do not have the disease as negative.

$$Specificity = \frac{TN}{(TN + FP)}$$

where,

TP – True Posotives, TN – True Negatives, FP – False Posotives, FN – False Negatives

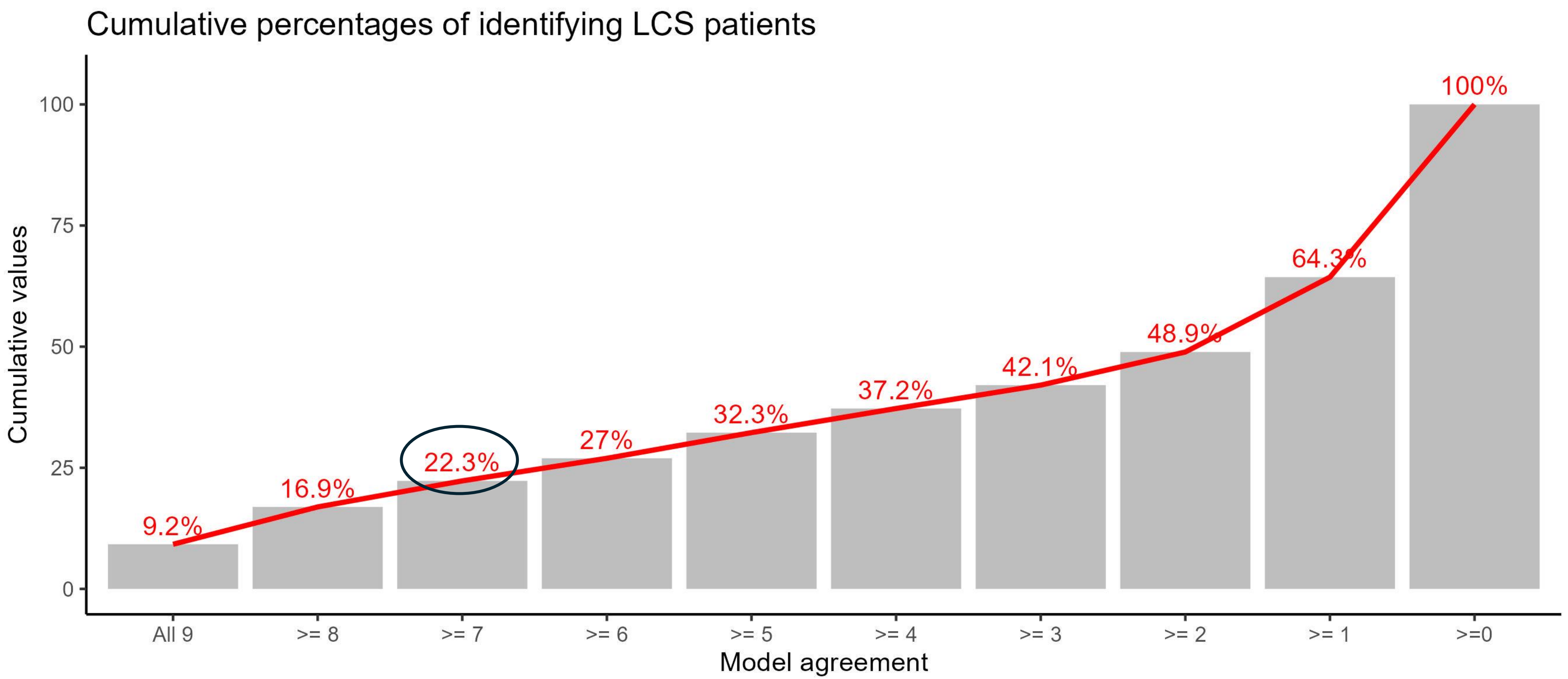
Model Results

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

- Most Important Features
  - Breathing/lung issues
  - Fatigue
  - Chest pain
  - Brain fog
  - Dizziness
  - Cough
  - Age group 70-79
- Noticeable Groups
  - Female (59.7%)
  - aged between 50 and 59 years old (18.8%)

Agreement between Nine Models



Conclusion

- Using natural language processing to identify 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 accuracy metrics: sensitivity of 0.95, specificity of 0.81, and AUC of 0.94.
- The LCS patient cohort created using this method is a valuable resource for conducting robust assessments of LCS clinical progression.

References

1. A. Nalbandian, 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, E. Y. Wan, Post-acute covid-19 syndrome, Nature medicine 27 (2021) 601–615.
2. A. Carf`ı, R. Bernabei, F. Landi, Persistent symptoms in patients after acute covid-19, JAMA : the journal of the American Medical Association 324 (2020) 603–605. doi:https://doi.org/10.1001/jama.2020.12603.
3. C. Huang, 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, B. Cao, 6-month consequences of covid-19 in patients discharged from hospital: a cohort study, The Lancet (British edition) 397 (2021) 220–232. doi:https://doi.org/10.1016/S0140-6736(20)32656-8.
4. doi:https://doi.org/10.1038/s41591-021-01283-z.S. Matharaarachchi, M. Domaratzki, A. Katz, S. Muthukumarana, Discovering long covid symptom patterns: Association rule mining and sentiment analysis in social media tweets, JMIR formative research 6 (2022) e37984–e37984. doi:https://doi.org/10.2196/37984.

Acknowledgement

- I would like to express my special thanks of gratitude to,
  - My supervisors, Dr. Saman Muthukumarana and Dr. Mike Domaratzki, for their excellent guidance.
  - Dr. Alan Katz for the constructive feedback.
  - The Manitoba Centre for Health Policy (MCHP) for providing the data.
  - The Department of Statistics and the staff for funding and resources.
  - My family and friends for their continuous support.

Thank You!