# Discovering long COVID symptom patterns: Association rule mining in social media tweets

Presented by: Surani Matharaarachchi

Dr. Saman Muthukumarana, Dr. Mike Domaratzki & Dr. Alan Katz

June, 01 2022

# Outline

# Introduction

- The COVID-19 pandemic
- “Long haulers”
- Long COVID: long-term health effects of COVID including Breathlessness, fatigue, and brain fog.
- Social media

# Long COVID

World Health Organization clinical case definition [3]:

"Long COVID-19 condition occurs in individuals with a history of probable or confirmed SARS CoV2 infection, usually three months from the onset of COVID-19, with symptoms that last for at least two months."

# Objective

To understand the patterns and behavior of long COVID symptoms, which is vital to improving our understanding of long COVID.

# Long COVID Symptoms

Table: List of Long COVID symptoms

| | Symptom | mayoclinic | NHS | CDC | WHO | [2] |
|---|---|---|---|---|---|---|
| 1 | Extreme tiredness (Fatigue) | ✓ | ✓ | ✓ | ✓ | ✓ |
| 2 | Shortness of breath or difficulty breathing | ✓ | ✓ | ✓ | ✓ | ✓ |
| 3 | Cough | ✓ | ✓ | ✓ | ✓ | ✓ |
| 4 | Joint pain | ✓ | ✓ | ✓ | ✓ | |
| 5 | Chest pain or tightness | ✓ | ✓ | ✓ | ✓ | ✓ |
| 6 | Problems with memory and concentration (“brain fog”) | ✓ | ✓ | ✓ | ✓ | ✓ |
| 7 | Difficulty sleeping (insomnia) | | ✓ | ✓ | ✓ | ✓ |
| 8 | Muscle pain | ✓ | | ✓ | ✓ | ✓ |
| 9 | Headache | ✓ | ✓ | ✓ | ✓ | ✓ |
| 10 | Fast or pounding heartbeat (heart palpitations)/ Tachycardia | ✓ | ✓ | ✓ | ✓ | ✓ |
| 11 | Loss of smell | ✓ | ✓ | ✓ | ✓ | ✓ |
| 12 | Loss of taste | ✓ | ✓ | ✓ | ✓ | ✓ |
| 13 | Depression or anxiety | ✓ | ✓ | | ✓ | |
| 14 | Fever | ✓ | | ✓ | ✓ | ✓ |
| 15 | Dizziness (lightheadedness) | ✓ | ✓ | ✓ | ✓ | ✓ |
| 16 | Worsened symptoms after physical or mental activities | ✓ | | ✓ | | |
| 17 | Pins-and-needles feeling | | ✓ | ✓ | ✓ | |
| 18 | Tinnitus, Earaches | | ✓ | | ✓ | ✓ |
| 19 | Diarrhoea | | ✓ | ✓ | ✓ | |
| 20 | Stomach aches | | ✓ | ✓ | | |
| 21 | Loss of appetite | | ✓ | | | ✓ |
| 22 | Sore throat | | ✓ | | | |
| 23 | Rash | | | ✓ | | |
| 24 | Mood changes | | | ✓ | | |
| 25 | Changes in menstrual period cycles | | | ✓ | ✓ | |
| 26 | Abdominal pain | | | | ✓ | ✓ |
| 27 | Neuralgias | | | | ✓ | |
| 28 | Allergies | | | | ✓ | |
| 29 | Body pain | | | | | ✓ |
| 30 | Nausea | | | | | ✓ |
| 31 | Weakness | | | | | ✓ |
| 32 | Numbness | | | | | ✓ |

# Data Collection & Pre-Processing

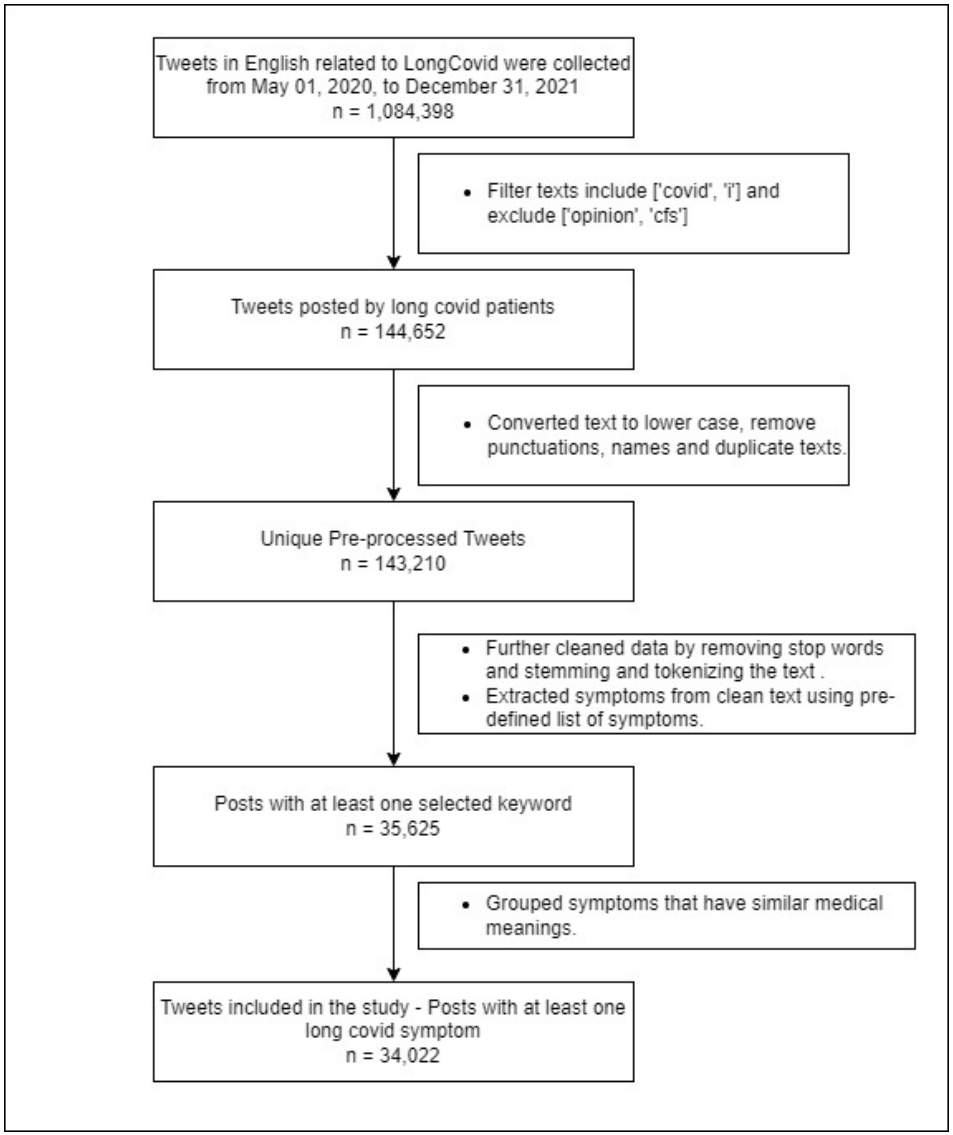
Data Collection:

- Worldwide long covid-related English tweets between 1 May 2020 and 31 December 2021.
- Used the Snscrape module in Python 3.8.
- Reduced the data set by limiting the population to those who suffered from COVID-19.

To ensure quality by developing a user-defined pre-processing function based on NLTK.
Pre-processing plan:

- Removed the hashtag symbol and its content, all non-English characters, repeated words and stop-words identified by NLTK, Special characters, punctuation, and numbers.

# Process


Tweets in English related to LongCovid were collected from May 01, 2020, to December 31, 2021
n = 1,084,398
Filter texts include ['covid', 'i'] and exclude ['opinion', 'cfs']
Tweets posted by long covid patients
n = 144,652
Converted text to lower case, remove punctuations, names and duplicate texts.
Unique Pre-processed Tweets
n = 143,210
Further cleaned data by removing stop words and stemming and tokenizing the text .
Extracted symptoms from clean text using pre-defined list of symptoms.
Posts with at least one selected keyword
n = 35,625
Grouped symptoms that have similar medical meanings.
Tweets included in the study - Posts with at least one long covid symptom
n = 34,022

# Sentiment Analysis

- Measure the sentiment expressed via Tweeter on long COVID.
- Sentiment analysis: a specific type of Natural Language Processing (NLP), computational linguistics, and text analysis.
- Classify the text into three classes: positive, negative, and neutral.
- Used the ‘TextBlob’ library: built on the NLTK library.

# Sentiment Analysis

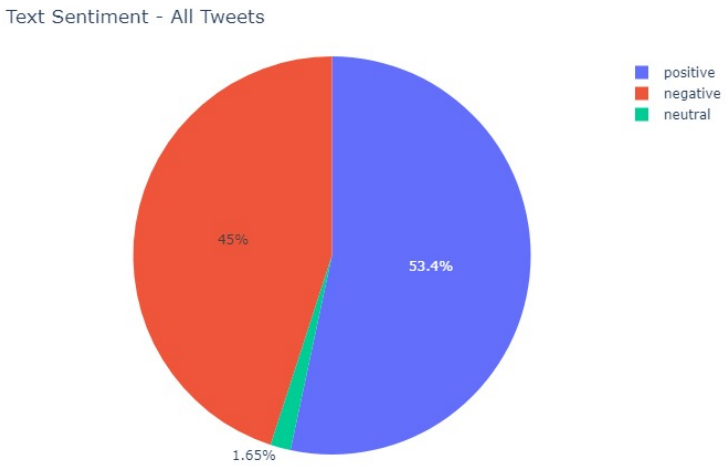



Figure: Classification of the sentiment score of all the posts

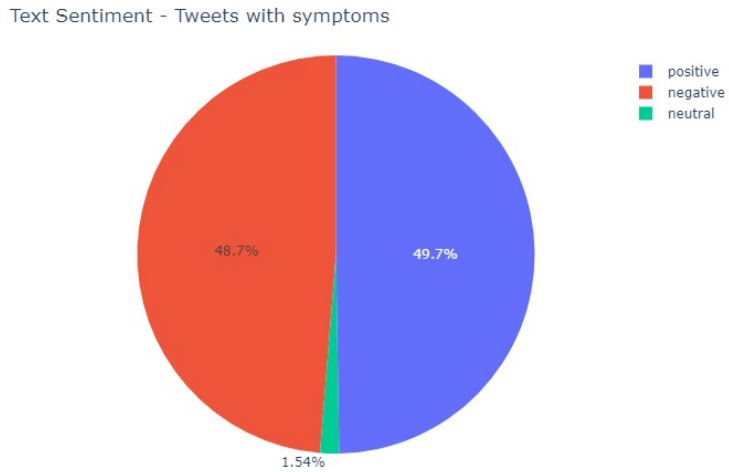
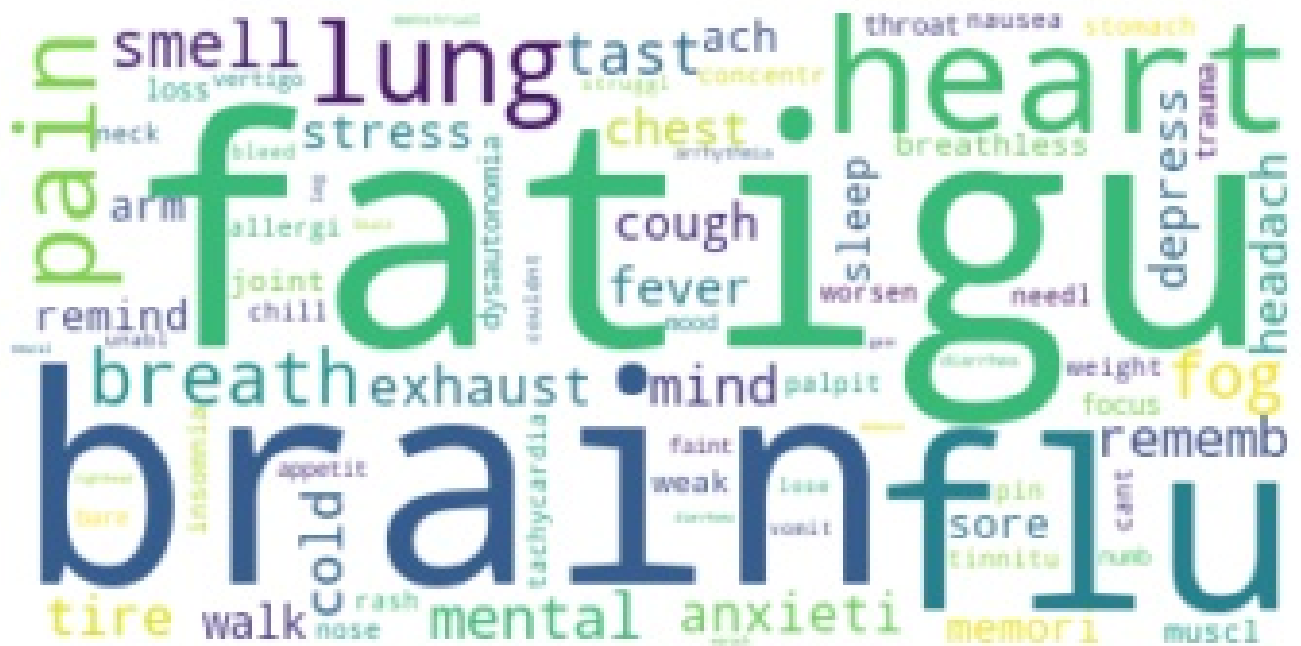



Figure: Classification of the sentiment score of the posts with at least one long COVID symptom.

# Word Extraction

created a data set for patients with symptom information using a pre-defined set of keywords.



# Collocation

- Symptoms often appear as more than one word in texts.
- Finding meaningful symptoms with only two words (bi-words).
- Collocation feature: reveals a phrase consisting of more than one word. These words more commonly co-occur in a given context than their individual word parts.
- Bigram-association measures:
  - Pointwise Mutual Information (PMI)
  - t-test with a frequency filter
  - Chi-square test

# Collocation Results

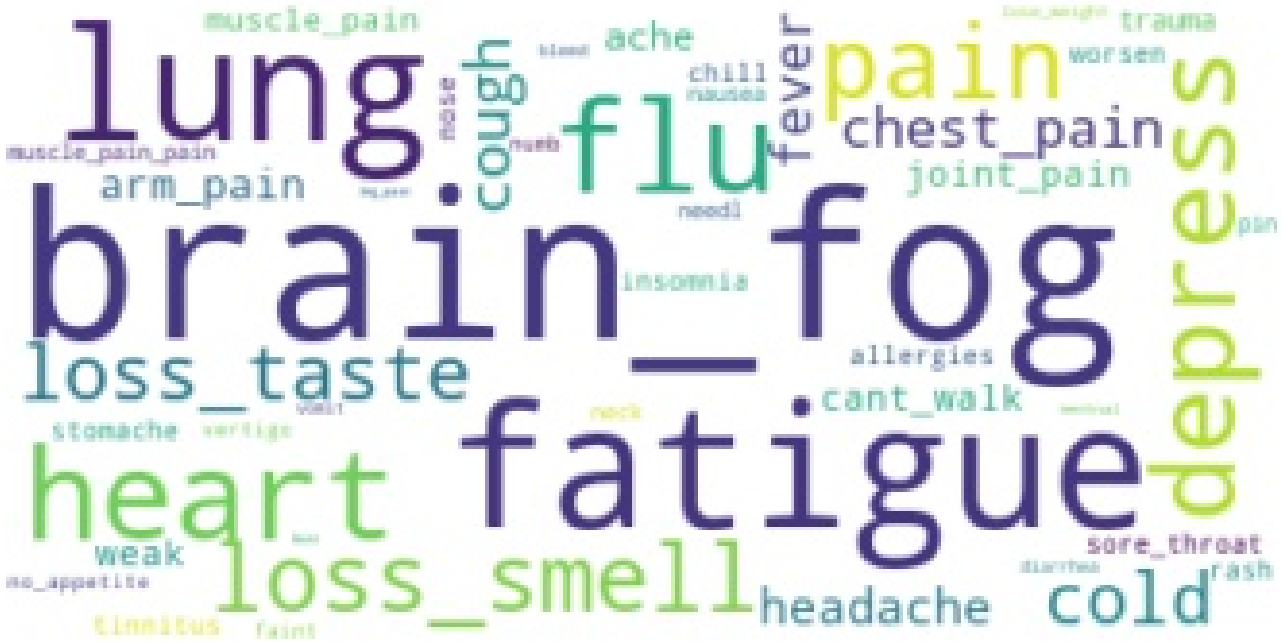
- We identified 20 such stemmed bi-words.
- Some bi-words have a similar medical meaning: we grouped similar words into categories for analysis.

Table: Pre-processed word corpus of stemmed symptoms

| Group | Symptoms |
|---|---|
| brain fog | 'brain fog', 'brain', 'fog', 'memori', 'mental', 'rememb', 'concentr', 'mind', 'remind', 'focus' |
| fatigue | 'fatigu', 'tire', 'exhaust' |
| lung | 'lung', 'breathless', 'breath' |
| cant walk | 'cant walk', 'struggl walk', 'unabl walk', 'couldnt walk', 'bare walk', 'unaid walk', 'stair walk' |
| depress | 'depress', 'mood', 'stress', 'anxieti' |
| lose weight | 'lose weight', 'loss weight' |
| insomnia | 'cant sleep', 'insomnia' |
| diarrhea | 'diarrhea', 'diarrhoea' |
| dizziness | 'dizz', 'lighthead' |
| heart | 'heart', 'heart palpit', 'tachycardia', 'dysautonomia', 'arrhythmia' |
| | 'headach', 'neck', 'arm', 'muscl pain', 'cough', 'chest pain', 'flu', 'joint pain', 'pain', 'rash', 'fever', 'loss smell', 'loss tast', 'cold', 'earach', 'vomit', 'chill', 'nausea', 'faint', 'gain weight', 'trauma', 'bodi', 'bleed', 'appetit', 'sore throat', 'pin needl', 'numb', 'tinnitu', 'buzz', 'hairfall', 'nose', 'stomach', 'menstrual', 'abdomin' |

# Reduced set of symptoms

The size parameter indicates the frequency of each symptom appearing in the tweets.



# Relative frequency of symptoms in long COVID patients.

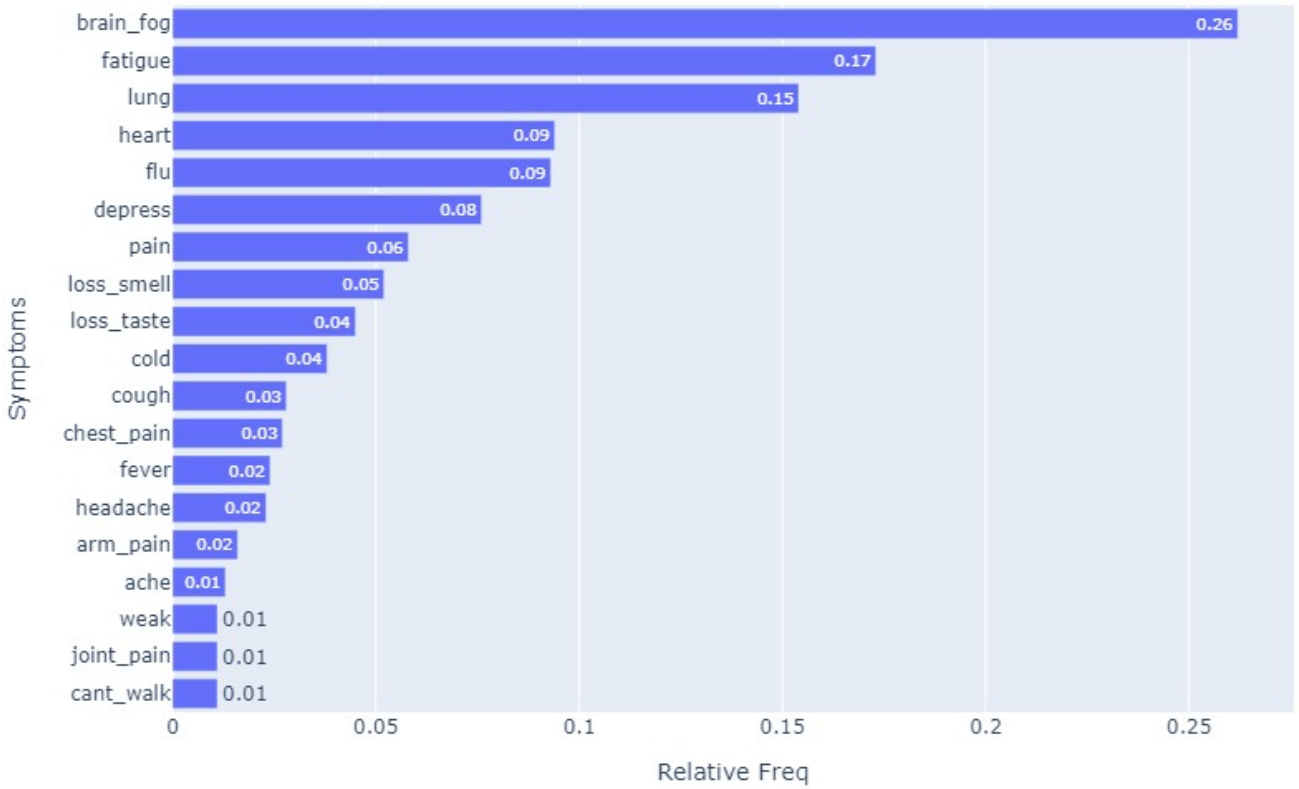



Figure: Relative frequency of symptoms in long COVID patients.

# Association Rule Mining

- An association: implications between different situations.
- Associations can be discovered and quantified using relational knowledge.
- Relational knowledge: identifies how entities are related and how entities and their relations are defined or described by models.
- These rules are called association rules, and the correlation analysis is known as association mining.

# Use of ARM

- To obtain insights into the poorly understood LCS by demonstrating the relationships and patterns among the symptoms described in the tweets.

- Automatically identifying new and useful symptom patterns in Long covid patient data using a most common Apriori rule-based data mining algorithm.

# Measures of Effectiveness of Rules

1. Support: Support indicates how frequently the item set appears in the data set.

$$supp(X \rightarrow Y) = supp(Y \rightarrow X) = P(X \cap Y)$$

2. Confidence: Confidence is the percentage of all transactions satisfying X that also satisfy Y.

$$conf(X \rightarrow Y) = P(Y|X) = \frac{P(X \cap Y)}{P(X)} = \frac{supp(X \rightarrow Y)}{supp(X)}$$

3. Lift: If the lift is > 1, that lets us know the degree to which those two occurrences are dependent on one another and makes those rules.

$$lift(X \rightarrow Y) = lift(Y \rightarrow X) = \frac{P(X \cap Y)}{P(X)P(Y)} = \frac{conf(X \rightarrow Y)}{supp(Y)} = \frac{conf(Y \rightarrow X)}{supp(X)}$$

# Symptom Rules

- Applied ARM to the symptom data considering one tweet as one transaction and identified symptom rules.

- Aimed to construct frequent item sets, having a user-specified threshold.
  - “confidence” threshold:0.1, minimum “support” threshold: above 0.001 and a “lift” greater than 1 for positively correlated rules.

# The highest confidence level-based detection

Table: Top 12 rules that have confidence $>$ 0.3.

| Rule | Antecedents | Consequents | Support | Confidence | Lift |
|---|---|---|---|---|---|
| R0 | (lung, loss_taste) | (loss_smell) | 0.0026 | 0.7739 | 14.9688 |
| R1 | (fatigue, loss_taste) | (loss_smell) | 0.0026 | 0.7395 | 14.3031 |
| R2 | (lung, loss_smell) | (loss_taste) | 0.0026 | 0.7063 | 15.8310 |
| R3 | (fatigue, loss_smell) | (loss_taste) | 0.0026 | 0.6377 | 14.2920 |
| R4 | (brain_fog, loss_taste) | (loss_smell) | 0.0019 | 0.6154 | 11.9026 |
| R5 | (loss_taste) | (loss_smell) | 0.0262 | 0.5870 | 11.3527 |
| R6 | (brain_fog, loss_smell) | (loss_taste) | 0.0019 | 0.5120 | 11.4751 |
| R7 | (loss_smell) | (loss_taste) | 0.0262 | 0.5065 | 11.3527 |
| R8 | (brain_fog, heart) | (lung) | 0.0034 | 0.3770 | 2.4439 |
| R9 | (heart, pain) | (lung) | 0.0011 | 0.3750 | 2.4306 |
| R10 | (ache) | (fatigue) | 0.0046 | 0.3421 | 1.9757 |
| R11 | (fatigue, cough) | (lung) | 0.0012 | 0.3158 | 2.0468 |
| R12 | (fatigue, heart) | (lung) | 0.0020 | 0.2991 | 1.9387 |

# Frame Title

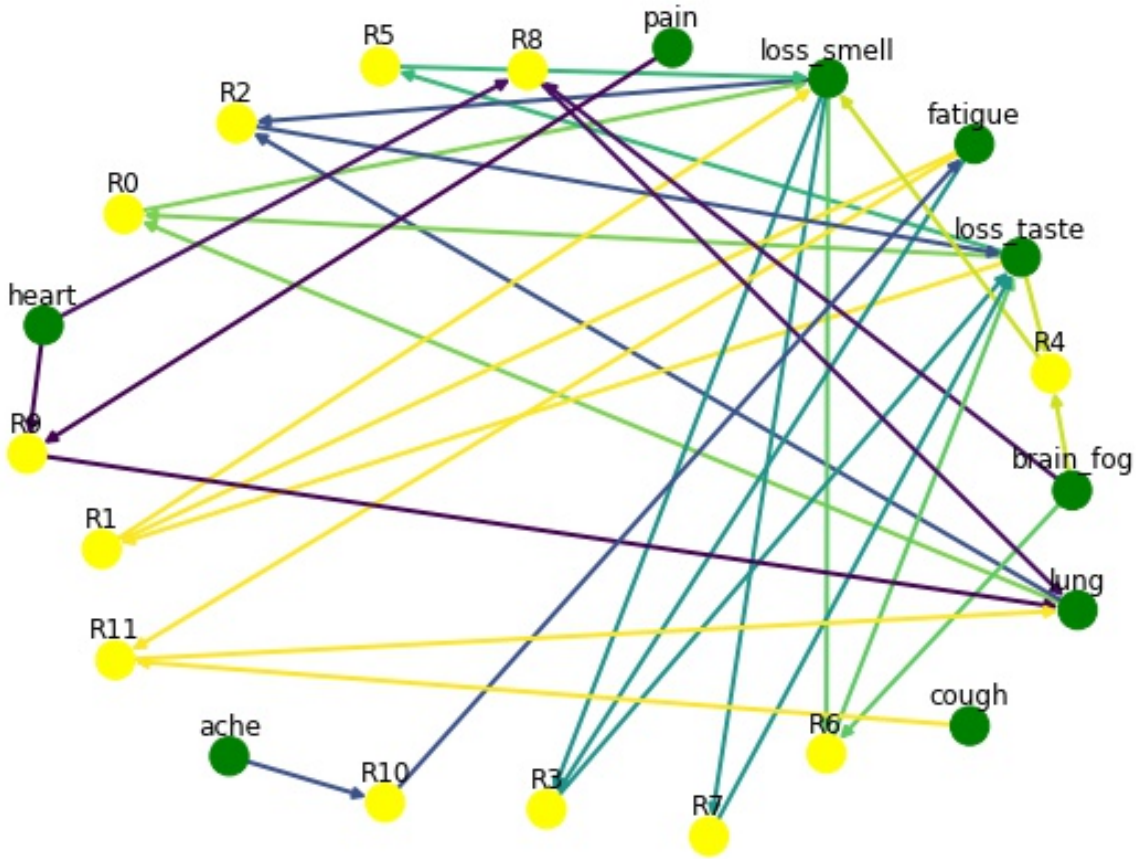



Figure: Association rules visualization.

# Discussion

- The symptoms associated LCS are still poorly understood.
- Analyzing social media conversations of long COVID-related patients allows one to understand the frequency and relationship between symptoms.
- First, we identified the symptoms and medical conditions related to long COVID. Then, we determined the patterns of symptoms and their associations.
- Brain fog, fatigue, and breathing/lung issues are the three most common symptoms identified by the analysis.

# Limitations

- Based on online twitter data with limited patient-level variables.
- No information about the demographics of tweeters was available.
- Only considered patients who shared their experiences with the public in English.
- Results can be affected by misinformation or false conversations on the Twitter platform.
- Have not investigated negatively correlated rules.

# Discussion Conti.

- Future research can build on this with clinical data sources such as EMR, adding individual covariates.
- Can also be further extended to detect and predict the consequences of a given set of symptoms using word popularity detection methods.
- Manuscript is submitted based on,
  - “Discovering long COVID symptom patterns: Association rule mining and sentiment analysis in social media tweets”.
- The pre-print is published [1],
  Matharaarachchi S, Domaratzki M, Katz A, Muthukumarana S, Discovering long COVID symptom patterns: Association rule mining and sentiment analysis in social media tweets. JMIR Preprints. 14/03/2022:37984
  https://doi.org/10.2196/preprints.37984

# References

[1] Matharaarachchi S, Domaratzki M, K. A. M. S. (2022). Discovering long covid symptom patterns: Association rule mining and sentiment analysis in social media tweets. *JMIR Preprints*, 37984.

[2] Singh, S. M. and C. Reddy (2020). An analysis of self-reported longcovid symptoms on twitter. *medRxiv*.

[3] Soriano, J. B., S. Murthy, J. C. Marshall, P. Relan, and J. V. Diaz (2021). A clinical case definition of post-covid-19 condition by a delphi consensus. *The Lancet infectious diseases*.

# Acknowledgment

I would like to express my special thanks of gratitude to,

- To my supervisors Dr. Saman Muthukumarana & Dr. Mike Domaratzki for their excellent guidance.
- To Dr. Alan Katz for providing the domain knowledge.
- To the department of Statistics and the staff for funding and resources.
- To my family and friends for the continuous support.

# Thank You!

email: matharas@myumanitoba.ca