Analyzing Long Covid-19 Data to Predict Long Covid-19 Cases

Katrina Dotzlaw, Ryan Dotzlaw, Sean Stzurm, Da Tan

Department of Computer Science

University of Manitoba

Winnipeg, MB, Canada

*Abstract*— Covid-19 data has been collected and analyzed since the start of the pandemic in 2019. Investigation of the post-Covid-19 condition (or Long Covid-19) began recently, and related data is being collected constantly. In this paper, we discover interesting associations in Long Covid-19 demographic data and cluster common symptoms. Using this information, we create a classifier that aims to predict the development of Long Covid-19 in patients. Our predictive model shows promise in identifying individuals at risk of developing Long Covid-19 and highlights demographic information that could indicate an increased risk of developing Long Covid-19.

Keywords—Covid-19, Long-Covid-19, data mining, prediction, association rules, clustering

# Introduction

Sean is writing

# Background and Related Works

## Long-Covid-19 Research

Sean is writing

## Related Works

Sean is writing

# Methodology

This section will describe the process of discovering association rules and creating a classifier.

## Association Rule Mining

Our approach to association rule mining applies the Apriori algorithm to the demographic data provided by the US Census Bureau [SITE] and identifies frequent characteristics. We use the following python libraries in our algorithm:

* math
* numpy
* matplotlib.pyplot: for graph creation
* mlxtend.frequent\_patterns: for the Apriori function and association\_rules function
* pandas: for data structures used with mlxtend

The census data is preprocessed before being read into our mining function. Certain demographic information included in the census (like employment status) will not be indicative of developing Long Covid-19 and is removed. Once the data has been cleaned, we read it into an array-like data frame. Additional processing must be done to the data before frequent patterns can be mined. We turn the numeric patient ages into a range and store the ranges in a new column, named ‘*age range’.* Next, we identify columns that contain categorical data and split them into multiple columns using binary mapping. For example, the column ‘*birth gender’* is categorical, containing [‘M’, ‘F’]. After applying the binary map, the categorical birth gender data becomes [[1,0], [0,1]], where 1 indicates the presence of a feature and 0 indicates the absence of a feature.

After preprocessing, the following columns will be used in subsequent frequent pattern mining and association rule mining:

* age range:
* symptom severity: none, mild, moderate, severe
* race: White, Black, Hispanic, Asian, other
* birth gender: male, female
* current gender: male, female, transgender, other
* vaccinated: yes/ no
* long covid: yes/no
* impacted: yes/ no
* booster: yes/no
* number doses:
* treat oral: received oral antiviral medication yes/no
* treat mono: received monoclonal antibody medication yes/no
* current symptoms: yes/no

Before mining frequent patterns, we generate graphs of demographic information with respect to Covid-19 and Long Covid-19. The occurrences in each column are counted and normalized to a percentage that is then displayed in a graph. Comparisons between graphs will be discussed in later sections.

Minimum support of a frequent item is determined based on the following formula:

[SITE]

Essentially, the minimum support is , where x increases as the number of rows in the dataset increases. Using the Apriori function from mlxtend, the calculated minimum support, and the columns previously mentioned, we identify frequent item sets.

The frequent item sets are then used in conjunction with the association rules function from mlxtend to mine interesting association rules. We determine rule to be interesting if it meets the minimum confidence of 0.3, where confidence of every rule is calculated by: [REF]. The association rules are then separated into two groups *long\_covid\_1* (where Long Covid-19 is the consequent) and *long\_covid\_0* (where **not** Long Covid-19 is the consequent). The former group of rules are sorted by confidence ascending.

## Predictive Model Creation

# Analysis

In this section, the association rules found are explained and the demographic information is analyzed. Additionally, the correctness of the classifier is examined, and predictive results are explained.

## Demographic Analysis – Week 46

The demographic information relating to Covid-19 and Long Covid-19 from the week 46 US Census Bureau [ref] was mined and graphed to compare distributions. The results of the demographic analysis will be compared to the results of demographic information collected and examined from a later census dataset.

We found that most Covid-19 patients are in the 30-60 age range. As shown in fig.1, the smallest percentage of individuals with Covid-19 are younger than 30 and older than 80.

Similarly to the age distribution of Covid-19 patients, most of the individuals experiencing Long Covid-19 are in the 30-60 age range (fig.2.). Individuals younger than 30 and older than 80 had the smallest incidence of Long Covid-19.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Fig. 1. Age Distribution of Covid-19 Patients Fig. 2. Age Distribution of Long Covid-19 Patients

In this dataset, 62.39% of individuals with Covid-19 were assigned female at birth, while 37.61% were assigned male at birth (fig.3). The amount of individuals assigned female at birth increases to 72.68% when examining individuals with Long Covid-19 (fig.4). The percentage of individuals assigned male at birth that have Long Covid-19 is 27.32%. This could indicate that individuals assigned female at birth are more likely to develop Long Covid-19, although this would require further analysis. It is unclear if this is an accurate representation of the birth gender of Covid-19 patients, or if there was just a higher instance of individuals assigned female at birth that responded to this survey.

Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generated

Fig. 3. Birth Gender Ratio for Covid-19 Patients Fig. 4. Birth Gender Ratio for Long Covid-19 Patients

Of the individuals with Covid-19 (fig.5), 60.91% identify as female while 36.79% identify as male. Only 0.41% identify as transgender and 1.89% identify as a different gender identity. The percentage of individuals who identify as female increases to 70.76% when examining individuals with Long Covid-19 (fig.6). The number of individuals identifying as male decreases to 26.63%. Transgender individuals and individuals that identify as a different gender identity make up 0.53% and 2.09% of the individuals with Long Covid-19. These results could indicate that more female-identifying individuals experience Long Covid-19.

Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generated

Fig. 5. Gender Identity of Covid-19 Patients Fig. 6. Gender Identity of Long Covid-19 Patients

The distribution of ethnicities is generally the same for both Covid-19 patients (fig.7) and Long Covid-19 patients (fig.8). White individuals represented the majority of both Covid-19 patients and Long Covid-19 patients with 73.5% and 71.15% respectively. The percentage of Hispanic individuals with Covid-19 who developed Long Covid-19 increased slightly from 10.75% to 12.79%. The percentage of Black individuals with Covid-19 who had Long Covid-19 also increased from 7.1% to 8.11%. Only 4.16% of individuals with Covid-19 in this dataset were Asian, and 3.09% of individuals with Long Covid-19 were Asian. The percentage of mixed individuals remained similar with 4.48% having Covid-19 and 4.86% having Long Covid-19.

Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generated

Fig. 7. Ethnicities of Covid-19 Patients Fig. 8. Ethnicities of Long Covid-19 Patients

The majority of Covid-19 patients in the week 46 dataset reported having either mild or moderate symptoms (fig.9), with 38.9% having mild symptoms and 41.33% having moderate symptoms. 14.48% of individuals with Covid-19 reported experiencing severe symptoms, while 5.3% reported having no symptoms. When examining symptom severity of Long Covid-19 patients, we found that 47.31% reported having moderate symptoms and 28.8% reported experiencing severe symptoms. This could indicate that Long Covid-19 symptoms are more likely to be moderate or severe, but further data analysis would need to be done. The number of individuals experiencing mild symptoms decreased to 22.4% when examining Long Covid-19 symptom severity. As seen in Fig.10., a small percentage of individuals reported experiencing no symptoms. These individuals could be asymptomatic and testing positive the amount of time required[[1]](#footnote-1) to receive a Long Covid-19 diagnosis, although it is more likely that they are part of the margin of error.

Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generated

Fig. 9. Symptom Severity of Covid-19 Patients Fig. 10. Symptom Severity of Long Covid-19 Patients

85.75% of the individuals with Covid-19 (fig. 11) in this dataset received at least 1 vaccine while 14.25% were not vaccinated. When examining individuals with Long Covid-19 (fig.12), we found that 83.85% of individuals had at least 1 vaccine, while 16.15% were not vaccinated.

Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generated

Fig. 11. Vaccination Rate of Covid-19 Patients Fig. 12. Vaccination Rate of Long Covid-19 Patients

## Explaining Interesting Association Rules – Week 46

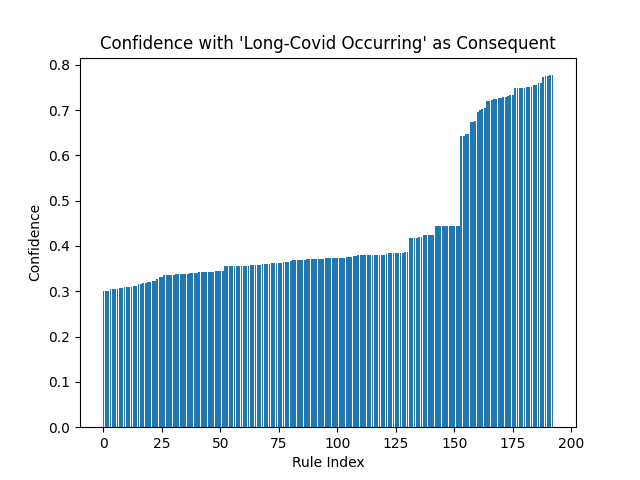


Fig. 13. Association Rules with ‘Long Covid Occurring’ as Consequent

Chart, histogram

Description automatically generated

Fig. 14. Association Rules with ‘Long Covid not Occurring’ as Consequent

## Demographic Analysis – Week 49

The demographic information from the week 49 dataset from the US Census Bureau [ref] will be examined and compared to the results of the week 46 analysis.

The age distribution of Covid-19 patients (fig. 15) for this dataset is similar to the distribution of the week 46 dataset (fig.1), although there is a sharper peak around age 40. The ages of the Long Covid-19 patients also had a sharper peak at 40 (fig. 16) but had a larger cluster of data in the 30-60 age range than in the Covid-19 data. The distribution of Long Covid-19 ages is similar to that of the week 46 data(fig.2).

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Fig. 15. Age Distribution of Covid-19 Patients Fig. 16. Age Distribution of Long Covid-19 Patients

According to the week 49 Covid-19 data analyzed, 55.89% of individuals were assigned female at birth. 44.11% were assigned male at birth. Compared to the week 46 data analyzed, the percentage of individuals assigned female at birth decreased, and the number of individuals assigned male at birth increased.

For the week 49 data, there is a higher percentage of individuals assigned female at birth who have Covid-19 (fig.17). Fig.18. shows that 67.43% of individuals assigned female at birth reported experiencing Long Covid-19. 32.57% of individuals assigned male at birth reported experiencing Long Covid-19. Since both datasets have a high population of individuals assigned female at birth experiencing Long Covid-19, it is possible that those individuals are more likely to develop Long Covid-19.

Chart, pie chart

Description automatically generated



Fig. 17. Birth Gender Ratio of Covid-19 Patients Fig. 18. . Birth Gender Ratio of Long Covid-19 Patients

When examining the gender identity of individuals with Covid-19, we found that 55.08% of patients currently identify as female and 43.46% currently identify as male (fig.19). Individuals who currently identify as transgender make up 0.47% of patients in this data set, while 1.0% of patients reported identifying as a different gender identity. This ratio indicates that there is higher percentage of female-identifying Covid-19 patients than male-identifying. Compared to the results of the analysis done on the week 46 data, there is a lower percentage of individuals who identify as female present in this dataset.

In our analysis of Long Covid-19 (fig.20), we found that the number of female-identifying individuals increased to 66.07% and the number of male-identifying individuals decreased to 31.83%. The number of transgender individuals and individuals that have a different gender identity also increased to 0.65% and 1.45% respectively. Similar to the results of the week 46 analysis, the increase in individuals that identify as female experiencing Long Covid-19 could indicate that female-identifying individuals are more likely to experience Long Covid-19.

Fig.19. Ratio of Gender Identity in Covid-19 Patients Fig.20. Ratio of Gender Identity in Long Covid-19 Patients



Chart, pie chart

Description automatically generated

According to our analysis on week 49 data, 76.45% of the Covid-19 patients are White. Hispanic individuals made up 9.52% of Covid-19 patients. 5.77% of individuals in this dataset are Black, while 3.95% are Asian. 4.3% of individuals with Covid-19 are mixed. Fig.21. shows that patients who are White make up most of the Covid-19 patients included in this dataset. Compared to the week 46 dataset, there is an increased number of White, Black, Hispanic, Asian, and mixed individuals in this dataset that experienced Covid-19.

Of the individuals reporting to have experienced Long Covid-19, 73.82% are White, 11.32% are Hispanic, 6.64% are Black, 5.73% are mixed, and 2.49% are Asian (Fig.22.). This could indicate that White individuals are more likely to experience Long Covid-19, however, recall that 76.45% of the Covid-19 patients in this data set are White and that demographic information can be skewed by inaccessible testing and treatment. Compared to the results from week 46, there is an increase in Black, Hispanic, Asian, and mixed individuals that develop Long Covid-19.





Fig.21. Ethnicity of Covid-19 Patients Fig.22. Ethnicity of Long Covid-19 Patients

The majority of Covid-19 patients reported experiencing mild or moderate symptoms, which is the same as the results of the week 46 analysis. As shown in Fig.23., 41.61% of individuals had mild symptoms and 41.77% had moderate symptoms. Severe symptoms – like hospitalization – were reported by 11.46% of Covid-19 patients in this data set. Only 5.16% of Covid-19 patients in this dataset reported experiencing no symptoms. Again, note that these results are from a self-reported survey. Symptom severity is relatively subjective and may not match the opinion of a health professional.

According to our analysis, symptom severity increased in individuals experiencing Long Covid-19, with the majority of individuals reporting moderate or severe symptoms. Since this matches the results from our previous analysis on week 46 data, it could be said that symptom severity increases when experiencing Long Covid-19. 49.46% of individuals reported experiencing moderate symptoms, while 25.45% of individuals reported experiencing severe symptoms. 23.47% of individuals with Long Covid-19 reported mild symptoms.





Fig.23. Symptom Severity in Covid-19 Patients Fig. 24. Symptom Severity of Long Covid-19 Patients

Among the Covid-19 patients in this dataset, 85.67% have received at least 1 vaccination, while 14.33% are unvaccinated (Fig. 25).

Similarly, 84.6% of individuals with Long Covid-19 reported to have received at least 1 vaccination. 15.4% of unvaccinated individuals reported experiencing Long Covid-19(Fig. 26.). Since these results follow the results of our previous analysis, it could indicate that Long Covid-19 is more likely to develop as a result of a breakthrough infection[[2]](#footnote-2), though this would need further research.





Fig.25. Vaccination Rate Among Covid-19 Patients Fig. 26. Vaccination Rate Among Long Covid-19 Patients

## Explaining Interesting Association Rules – Week 49

In our dataset, 27.56% of individuals reported experiencing Long Covid-19, while 72.44% reported not experiencing it. We are mainly interested in associations concerning Long Covid-19, although rules with ‘*long covid not occurring’* are also found.

An association rule with ‘*Long Covid Occurring’* as the consequent is of the form: .

An association rule with ‘*Long Covid Not Occurring’* as the consequent is of the form:

Comparing the graphs with ‘*Long Covid Occurring’* (fig. 27.)and ‘*Long Covid Not Occurring’* (fig. 28.) in the consequent, we can see that there are more rules in the latter group. Since 72.44% of patients in this dataset reported not experiencing Long Covid-19, it makes sense that we found more associations with ‘*Long Covid Not Occurring’* in the consequent*.*



Fig. 27. Confidence of Rules with ‘Long Covid Occurring’ in the Consequent

Chart, histogram

Description automatically generated

Fig. 28. Confidence of Rules with ‘Long Covid Not Occurring’ in the Consequent

Over 400 rules were found with confidence ranging from approximately 0.6 to greater than 0.8. The found rules were sorted in ascending order by confidence.

Our association rule mining discovered approximately 33 interesting associations with ‘*Long Covid Occurring’* in the consequent. The rules are sorted by confidence ascending. The confidence of all interesting rules ranges from approximately 0.3 to approximately 0.8.

[DISCUSS TOP X ASSOCIATION RULES FOR WEEK 49]

## Evaluating Model Correctness

# Conclusions

## Conclusions

## Limitations

Since Long Covid-19 is an active area of research, it was difficult to find open-source data to analyze. Most of the data sets we found required credentials from a reputable institution. Additionally, some of the data sets we originally planned to analyze were removed, presumably for private use by the CDC or governments. Due to the removal of data sets, we were unable to examine connections between Covid-19 variants and Long Covid-19 diagnosis.

Almost all the data sets we analyzed were self-reported surveys, which can often be biased and exclusionary. Individuals experiencing Covid-19 and Long Covid-19 that do not have internet access would likely be excluded from such surveys. Individuals that do not have access to safe and reliable health care might report symptoms inaccurately.

We were unable to find a publicly accessible data set containing Covid-19 symptoms and Long Covid-19 symptoms to develop a symptom-based predictive model. Again, this is likely because Long Covid-19 is an active area of research.

## Future Work

Ideally, Long Covid-19 data will become publicly available as research into the virus progresses. Finding supervised data that contains both Covid-19 and Long Covid-19 data would allow the creation of a symptom-based classifier.

Additionally, Covid-19 variant data could be used to find associations between different variants and the development of Long Covid-19. Our predictive model could be expanded to predict the development of Long Covid-19 based on variant diagnosis.

Further investigation could be done into the relationship between vaccination status and the development of Long Covid-19, as our demographic analysis indicated that Long Covid-19 could arise as a result of a breakthrough infection.

##### Acknowledgment

##### References

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science,

1. The required amount of time symptoms persist after initial Covid-19 diagnosis is debated among health organizations. [↑](#footnote-ref-1)
2. An infection of vaccinated individuals. [↑](#footnote-ref-2)