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

## Explaining Interesting Association Rules – Week 46

## Demographic Analysis – Week 49

As described in the *methodology* section, we graphed demographic information relating to Covid-19 and Long Covid-19 to compare distributions.

Using the week 49 data set from the US Census Bureau [reference], 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.

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

According to the US census data analyzed, 55.89% of individuals were assigned female at birth. 44.11% were assigned male at birth. As shown in in fig. 3, there is a higher percentage of individuals assigned female at birth who have Covid-19. 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.

Fig. 4 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. This could indicate that individuals assigned female at birth are more likely to develop Long Covid-19 but note that the data set used contains a majority of individuals assigned female at birth.

Chart, pie chart

Description automatically generated



Fig. 3. Birth Gender Ratio of Covid-19 Patients Fig. 4. 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. 5). 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.

Stuff about long covid gender identity here.

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



Missing graph

In the US census survey, individuals identified their ethnicity based on 5 categories: White, Hispanic, Black, Asian, and mixed. 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.7. shows that patients who are White make up most of the Covid-19 patients included in this dataset.

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. 8.). 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.





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

The majority of Covid-19 patients reported experiencing mild or moderate symptoms. As shown in Fig.9, 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. 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. 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.





Fig. 9. Symptom Severity in Covid-19 Patients Fig. 10. 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. 11).

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. 12.). This 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. 11. Vaccination Rate Among Covid-19 Patients Fig. 12. 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. 13.)and ‘*Long Covid Not Occurring’* (fig. 14.) 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. 13. Confidence of Rules with ‘Long Covid Occurring’ in the Consequent

Chart, histogram

Description automatically generated

Fig. 14. 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 (fig. 15.) 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)