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
* race
* birth gender
* current gender
* vaccinated
* long covid
* impacted
* booster
* number doses
* treat oral
* treat mono
* current symptoms

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

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

1. Age Distribution of Covid-19 Patients

Chart, scatter chart

Description automatically generated

1. Age Distribution of Long Covid-19 Patients

Similarly, most individuals with Long Covid-19 are in the 30-60 age range. As shown in Fig.2, individuals younger than 30 and older than 80 have the smallest incidence of Long Covid-19.

Chart, pie chart

Description automatically generated

Fig. 3. Birth Gender Ratio of 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. Birth Gender Ratio of Long Covid-19 Patients

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.

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.

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



[Grab your reader’s attention with a great quote from the document or use this space to emphasize a key point. To place this text box anywhere on the page, just drag it.]

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

Stuff about long covid gender identity here.



Fig. 7. Ethnicity of Covid-19 Patients

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.



Fig. 8. Ethnicity of Long Covid-19 Patients

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. 9. Symptom Severity in 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.



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

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. 11. Vaccination Rate Among Covid-19 Patients



Fig. 12. Vaccination Rate Among 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.

## Explaining Interesting Association Rules

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:



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

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*.* 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.





Fig. 15. Association Rules with ‘Long Covid Occurring’ in the

Consequent, with Support and Confidence

## Symptoms prevalence analysis

Da Tan: please polish the language (where should this part goes?)

Next, we explored the relations between demographic groups and Long Covid-19 symptoms-combinations and reported the symptoms that have significantly different occurrence-frequencies among the groups. This was done by using the Kenya, Malawi Long Covid-19 effect survey dataset [[Kenya, Malawi, Long Covid-19 effects survey dataset - Humanitarian Data Exchange (humdata.org)](https://data.humdata.org/dataset/long-covidresearchagenda)].

The survey consists of 677 and 679 Long Covid-19 cases from Kenya and Malawi respectively. 6 demographic features, as well as 35 post-Covid symptoms, were collected. For all the demographic groups, the frequencies of the symptoms were reported, as shown in Figure 16, with headache, cough, and fatigue as the most prevalent post-Covid symptoms.

Chart, bar chart

Description automatically generated

Fig. 16. Rank of Long Covid-19 symptoms occurrences

We also discovered that there are significant differences in symptom occurrences among the demographic subgroups, as demonstrated in Figure 17. Specifically, for each demographic feature, we calculated the frequencies for each group, performed a chi-square test and ranked the symptoms that are significantly different among the subgroup (p-value threshold = 0.05). Our analysis showed that Long Covid-19 patients of 50 years old or elder experienced more symptoms such as non-communicable diseases, pre-existing conditions, fatigue, and joint pain, etc. It is reasonable because elder people are biased to have such conditions, and it is hard to link this prevalence with Long Covid-19 since the lack of control datasets that are Long-Covid-19-negative. We also noticed that in this dataset, more women experienced symptoms such as shortness of breath, chest pain, etc. Moreover, the prevalence of many symptoms such as cough, running nose, recurrent fever, etc. was significantly different between the two countries Kenya and Malawi. We are not able to trace the reason for this bias, which may be due to the different Covid-19 subtypes in these two countries, or the time distance of conducting the two surveys. In addition, it is noticed that there was a statistical difference in symptom prevalence between the employed and unemployed groups, and the unemployed group tends to have a higher chance to develop symptoms such as fatigue, cough, etc. This might be because unemployed people averagely tend to be elderly or have worse health conditions than the employed population. This is evidenced by the fact that pre-existing conditions and non-communicable diseases are much more prevalent in the unemployed population. There are other demographic differences in features of living type and the number of people living with them, but we still lack evidence to trace the cause of such deviations.

Chart, bar chart, waterfall chart

Description automatically generated

Fig. 17. Symptoms occurrences differences among demographic groups

Next, we focused on the age-gender subgroups and compared the similarities between them in terms of the representative symptoms prevailing in each group. We did not consider other features such as employed, country, living types and the number of people living, because we deemed those social-state features irrelevant to our study. Also, we reduced the number of symptoms by eliminating those with less than 5% of overall occurrence, since infrequent symptoms increase the dimensionality for the similarity comparison as well as the clustering algorithm.

For each of the 6 age-gender subgroups, we calculated the occurrence percentages and computed the similarities among the 6 groups. Specifically, for each pair of age-gender subgroups and , we computed the cosine similarity between them:

The higher the value, the more similar the two age-gender groups. As shown in Table 1 and Figure 18, there is a significant dissimilarity between the group of “age > 50” and “age <= 50”, and there are no obvious differences between the genders as well as the two younger age groups.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 18-33\_Male | 34-50\_Female | 34-50\_Male | 18-33\_Female | >50\_Male | >50\_Female |
| 18-33\_Male | 1 | 0.976 | 0.976 | 0.985 | 0.915 | 0.887 |
| 34-50\_Female | 0.976 | 1 | 0.991 | 0.966 | 0.964 | 0.96 |
| 34-50\_Male | 0.976 | 0.991 | 1 | 0.96 | 0.962 | 0.948 |
| 18-33\_Female | 0.985 | 0.966 | 0.96 | 1 | 0.889 | 0.871 |
| >50\_Male | 0.915 | 0.964 | 0.962 | 0.889 | 1 | 0.975 |
| >50\_Female | 0.887 | 0.96 | 0.948 | 0.871 | 0.975 | 1 |

Table1. Cosine Similarities among the age-gender groups

Square

Description automatically generated with low confidence

Fig. 18. Heatmap showing the similarities among the age-gender groups

## Unsupervised Learning

To explore the relationship between the inner structure of the symptoms and the age-gender demographics, we performed K-mode clustering on the Kenya and Malawi survey datasets. With the reduced 15 symptoms, the optimized number of clusters is 11 according to the metric of Total within-clusters Sum of Squares, as shown in Fig. 19. The elbow point usually indicates an optimal K value, and we chose K=11 at the elbow point, and this, on the other hand, avoids for too many clusters. Table 2 demonstrates the 11 modes representing the 11 clusters. Then we mapped the 11 modes to the original age subgroups, as shown in Table 3. No obvious overlapping was found between the clustering result and the age groups. This might be because most of the symptoms occurred in a low frequency, and there are few symptoms (as shown in Figure 17) that occur with significant different frequencies between the age groups. However, we could see some high concentration for the “>50” group in cluster 3, the one with mode of “non-communication diseases” and “pre-existing conditions”, which are obviously much more prevalence in the “>50” groups.

A screenshot of a game

Description automatically generated with medium confidence

Table 2. The 11 modes for the K-mode clustering

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster/ age group | 18-33 | 34-50 | >50 |
| 1 | 90 | 80 | 11 |
| 2 | 1 | 24 | 10 |
| 3 | 8 | 34 | 21 |
| 4 | 26 | 23 | 0 |
| 5 | 82 | 44 | 7 |
| 6 | 17 | 50 | 9 |
| 7 | 157 | 129 | 14 |
| 8 | 67 | 47 | 8 |
| 9 | 186 | 131 | 16 |
| 10 | 25 | 24 | 1 |
| 11 | 3 | 8 | 3 |

Table 3. Mapping of the 11 clusters to the 3 age groups

Chart, line chart

Description automatically generated

Fig.19. The optimization path for the number of clusters

## Supervised Learning

In order to explore how well the features in the US Census dataset can predict whether a Covid-19 patient will develop Long Covid-19, we developed a predictive classifier on it. Since the model is predictive, we removed the features “Current Symptoms” and “Impacted”, which are only available when the instance develops Long Covid-19. Then we removed the cases that did not have Covid-19 and those whose “Symptom Severity” feature is NA (taking up less than 1%). We further removed “Treat Oral” and “Treat Mono” because more than 90% of the values of those two features are NAs. The above-preprocessing step resulted in 23349 cases with 7 predictive variables. Among the 23349 cases that had Covid-19, 6468 of them developed Long Covid-19 and 16881 didn’t. Since most of the classification algorithms are sensitive to unbalance datasets, we randomly sampled 70% (4528) of the positive cases and took them in the training dataset, and then randomly sampled an equal number of negative cases and put them into the training set. And all the remaining cases were the independent test set, which comprised 1940 positive and 12353 negative cases.

Before training the model, descriptive analysis and feature selection were performed. Figure 20 illustrates the seven candidate predictors, among which, only the feature “Age” is numerical, and the other 6 features are categorical. Intuitively from Figure 20, we noticed that only the feature “Symptom Severity” seems to be a strong distinguisher.

The feature selection process was performed by using “Boruta\_8.0.0” package of R 4.2.0. Boruta selects features by wrapping the Random Forest algorithm inside, and then randomly shuffles, trains the data, and reports importance ranks for the features [8]. Among the 7 candidate predictors, “Vaccinated” was rejected by Boruta, “Age” was tentative, while the other predictors were confirmed. Among those, “Symptom Severity” was reported to have the highest importance rank, which agrees with our descriptive analysis.

Chart, bar chart

Description automatically generated

Fig 20. Descriptive analysis for the US Census dataset

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | meanImp | medianImp | minImp | maxImp | normHits | decision |
| AGE | 2.731 | 2.886 | -3.977 | 7.797 | 0.424 | Tentative |
| RACE | 7.793 | 7.724 | 3.293 | 13.994 | 0.96 | Confirmed |
| BIRTH\_GENDER | 24.65 | 24.551 | 16.532 | 30.367 | 1 | Confirmed |
| VACCINATED | 0.013 | -0.449 | -4.021 | 3.907 | 0.01 | Rejected |
| NUMBER\_DOSES | 7.922 | 7.951 | 3.177 | 13.942 | 0.96 | Confirmed |
| BOOSTER | 14.005 | 14.124 | 5.014 | 21.78 | 1 | Confirmed |
| SYMPTOM\_SEVERITY | 111.986 | 112.556 | 96.694 | 123.554 | 1 | Confirmed |

Table 4. Feature selection result from Boruta algorithm

Next, we removed the feature “Vaccinated”, as rejected by Boruta, and trained classifiers on the processed data. We built two classifiers with Decision Tree and Random Forest algorithms respectively. Then we evaluated the model performances on the independent test set with metric of AUC (Area Under Curve).

First, we built a Decision Tree using the R package “rpart”. We used the R package “caret” to tune the parameters “cp” (complexity of the tree) and “max\_depth” (maximal depth that the tree can grow). 10-fold cross validation was used in the tuning process and the optimization paths are demonstrated in Figure 21. With the optimized parameters “cp=0.003” and “maxdepth=3”, we constructed a decision tree and the tested the model on the test dataset. The AUC for the model performance is 0.706 (±0.011), as shown in Figure 21. The final tree is demonstrated in Figure 22.

Next, we constructed a Random Forest model using the R package “randomForest”. We tuned the parameter “mtry” of the model (number of features for growing a random tree) by the “tuneRF” function. Then with the optimized “mtry=2”, we trained a Random Forest model, and evaluated the model on the independent test dataset. The AUC for the Random Forest model is 0.721 (±0.011), as shown in Figure 22. We also reported the feature importance rank returned by the Random Forest model, as demonstrated by Table 5.

Chart

Description automatically generated with low confidence

Fig. 21. The optimization paths for parameters of the decision tree. The tuning metric is AUC, and the parameters tuned are complexity and maximal depth. Subgraphs at the second line indicate the performances of the two models Decision Tree and Random Forest, in terms of AUC.

Timeline

Description automatically generated

Fig. 22. The decision tree model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | no | yes | MeanDecreaseAccuracy | MeanDecreaseGini |
| AGE | -0.002 | 0.003 | 0.001 | 198.178 |
| RACE | 0.004 | 0.004 | 0.004 | 75.789 |
| BIRTH\_GENDER | 0.008 | 0.023 | 0.015 | 79.842 |
| NUMBER\_DOSES | 0.009 | -0.002 | 0.004 | 38.779 |
| BOOSTER | 0.015 | 0.001 | 0.008 | 36.166 |
| SYMPTOM\_SEVERITY | 0.091 | 0.1 | 0.095 | 472.718 |

Table 5. The feature importance rank by the Random Forest model

The feature importance rank shown in Table 5 shows that the most important feature for the RF classifier is “Symptom\_severity”, followed by “Age”. This agrees with our descriptive analysis as shown in Figure 20.

Although we don’t have many good predictors as shown in our descriptive analysis, the performance of the two classifiers is fine, with AUCs 0.706 and 0.721 for the Decision Tree and Random Forest classifier, respectively. Random Forest generally perform better than the Decision Trees, as evidenced in our study. Besides the metric of AUC, we also reported other model evaluation metrics such as Accuracy, F1, etc., as shown in Table 6. We used threshold = 0.5 to assign the predictions to the two classes. To use which performance metric is dependent on the practical application of the model.

|  |  |  |
| --- | --- | --- |
| metric/model | Decision Tree | Random Forest |
| AUC | 0.706 | **0.721** |
| Accuracy | **0.711** | 0.676 |
| Sensitivity | **0.723** | 0.678 |
| Specificity | 0.607 | **0.668** |
| Precision | 0.922 | **0.929** |
| Recall | **0.727** | 0.678 |
| F1 | **0.813** | 0.784 |

Table 6. The performance metrics of the Decision Tree and Random Forest models

## 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,

[8] Kursa M B, Rudnicki W R. Feature selection with the Boruta package[J]. Journal of statistical software, 2010, 36: 1-13.

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)