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 analyzed, 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. The smallest percentages of individuals with Covid-19 are younger than 30 and older than 80, as seen in Fig. 1.

Chart, scatter chart

Description automatically generated

## Explaining Interesting Association Rules

[THIS IS FOR DEMO ANALYSIS – move later]

## 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.

##### 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. Age distribution of 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, scatter chart

Description automatically generated

1. Age distribution of Long Covid-19 patients