**1.0 Introduction**

The Coronavirus disease 2019, also known as COVID-19, was first discovered in Wuhan City, China during December 2019 [1]. Since then, COVID-19 has spread worldwide (1) and according to the World Health Organization, as of December 2022 nearly 650,000,000 individuals have been diagnosed with COVID-19 (2) and over 6,600,000 individuals have died after being diagnosed (2). While symptoms can vary between individuals, the World Health Organization has listed fever, fatigue and cough, along with sore throat, shortness of breath, headache and chest pain as common symptoms (3).

Some individuals experience COVID-19 symptoms once they no longer test positive for the disease (12). Consensus on a definition of Post COVID-19 condition, also known as Long Covid (Reference), has not yet been reached. The National Institute for Health and Care Excellence defines Long Covid as “signs and symptoms that develop during or after an infection consistent with COVID-19, continue for more than 12 weeks and are not explained by an alternative diagnosis” (4). They continue by saying Long Covid “usually presents with clusters of symptoms, often overlapping, which can fluctuate and change over time and affect any system in the body” (4). According to the World Health Organization, a potential definition of Long Covid could be new or ongoing symptoms occurring after the initial four-week infection period with symptoms including any of the discovered COVID-19 symptoms (World Health Organization, 2022). According to [6], “any patient with covid-19 may develop long covid, regardless of the severity of their infection and the intensity of the treatment they received” [6]. Vu and McGill have classified Long Covid into two stages: “Ongoing symptomatic COVID-19” (Reference) which they define as symptoms occurring between 4 and 12 weeks after an individual has been infected with Covid-19 (Vu & McGill, 2021) and “Post Covid-19 Syndrome” (Reference) which they define as symptoms occurring 12 weeks or more after an individual has been infected with COVID-19 (Vu & McGill, 2021). The World Health Organization estimates that up to 20% of individuals diagnosed with COVID-19 will develop Long Covid (World Health Organization, 2022), with (15) reporting nearly 40% of the individuals in their study having symptoms for longer than one month and nearly 15% having symptoms for more than three months (15). Early in the pandemic, it was reported that nearly 90% of individuals diagnosed with COVID-19 who were hospitalized at a particular hospital in Rome experienced Long Covid [12]. There is some evidence stating that Long Covid may be more common in women [11] and that symptoms of Long Covid can depend on patient age as [11] states “we do seem to be getting different symptom clusters in different ages” [11].

The world is still in the midst of the COVID-19 pandemic. According to the Centers for Disease Control and Prevention, there were nearly 470,000 new COVID-19 cases in the United States during the week of December 7th to 14th, 2022 [5]. There has been a push to report on Long Covid similar to how regular COVID-19 cases are tracked and reported [11]. Since a material proportion of individuals who have been diagnosed with COVID-19 will be diagnosed with Long Covid (World Health Organization, 2022), we believe that identifying relationships and patterns between Covid-19 symptoms and Long Covid development could help contribute to a better understanding of the illness and allow for more effective treatment through monitoring individuals most at risk for developing Long Covid.

* 1. **Our Contribution**

Our group is attempting to create a classifier to predict the development of Long Covid in individuals who are diagnosed with COVID-19. We begin by using the Apriori Algorithm to mine interesting associations occurring in individuals that would imply the development of Long Covid using data from the United States Census Bureau (14). We are then using K-Mode Clustering to explore datasets containing demographic and symptom information from Long Covid patients in order to train the classifier. Finally, we are analyzing the correctness of the classifier and providing an explanation of our results, as well as discussing limitations of our analysis.

Our research includes datasets containing Long Covid data. The first dataset used in our analysis was data from surveys conducted by the United Nations Office for the Coordination of Humanitarian Affairs in Kenya and Malawi detailing demographical information and symptoms of individuals who have been diagnosed with Long Covid in those countries (13). These surveys included patient information such as gender, age, employment status, the number of other people the individual is living with, whether the individual suffered from any pre-existing conditions and any COVID-19 symptoms experienced. The next dataset important to our analysis is the from the United States Census Bureau containing household survey data which includes a question about whether the individual surveyed experienced Long Covid (14).

1: https://www.sciencedirect.com/science/article/pii/S2319417020300445

2: <https://covid19.who.int/>

3: <https://www.who.int/health-topics/coronavirus#tab=tab_3>  
4: <https://www.nice.org.uk/guidance/ng188>  
5: <https://covid.cdc.gov/covid-data-tracker/#trends_weeklycases_select_00>

6: <https://www.bmj.com/content/374/bmj.n1648.full>

7: <https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-021-11013-2>

8: <https://www.medrxiv.org/content/10.1101/2021.03.08.21253124v1.full-text#T1>

9: <https://erj.ersjournals.com/content/early/2022/08/04/13993003.00970-2022>

10: <https://doi.org/10.1038/s41591-021-01292-y>

11: <https://www.bmj.com/content/370/bmj.m3489>

12: <https://www.bmj.com/content/370/bmj.m2815.long>  
13: <https://data.humdata.org/dataset/long-covidresearchagenda>  
14: <https://www.census.gov/programs-surveys/household-pulse-survey/datasets.html>

15: <https://www.medrxiv.org/content/10.1101/2020.10.07.20208702v3>