**Impacts of Seasonal Change on SARS-CoV-2 Infection Rates**

BIO812 - Group Assignment - Written Report

Courtney Lanigan (203081690), Andrew Le (20314759), Kyla Tozer (20042692), Sanjana Venkatesh (20308106)

# - These authors contributed equally to this work

**Introduction**

The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) emerged in December of 2019, in Hubei province, China [1]. January 12th, 2020 the virus was isolated and sequenced, identifying the virus as the seventh member of the *Coronavirindea* family [1]. From December to March of 2020, the virus spread rapidly around the world and on March 11th, 2020, the World Health Organization declared a global pandemic for this virus. As of April 14th, 2022, there have been over 500 million cases spanning in 224 countries and territories, resulting in 6-million deaths [2]. Of the total cases, 99% of infections have resulted in mild to moderate illness, while 1% lead to hospitalization [2]. SARS-CoV-2 is the seventh viral member of the *Coronavirinae* subfamily infecting humans [1]. *Coronavirinae* viruses are divided into four genera which include: *Alphacoronavirus, Betacoronavirus, Gammacoronavirus,* and *Deltacoronavirus.* Alpha- and Beta- genera infect humans, whereas Gamma- and Delta- infect primarily birds [1]. All lineages of Coronaviruses (CoVs) are sorted based on their phytogenic clustering [3]. The *Alphacoronaviruse*s include HCoV-NL63 and HCoV-229E; while, *Betacoronavirus’*, include HCoV-OC43, HCoV-HKU1, Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV), Middle East Respiratory Syndrome Coronavirus (MERS-CoV), and the newly identified SARS-CoV-23.

**Viral Entry and Manifestation**

SARS-CoV-2 targets epithelial cells in the respiratory tract, specifically cells expressing the angiotensin converting enzyme-2 receptor (ACE2), which is present ubiquitously around the human body [4]. Infection of SARS-CoV-2 begins with binding of the virus to an ACE2 receptor using the S-protein.

There are critical features and indications which impact the spread and severity of a virus and these features include: pathophysiology, immune response, tropism, viral load, viral shedding, molecular mechanisms, and binding affinity [5,6]. SARS-CoV-2 does not have a linear response profile and varies from individual to individual. Generally, an infected patient develops symptoms within 5-6 days following viral exposure, this time period is known as the incubation period. Within the first 7-days post viral exposure, viral load is at its highest, followed by a steady decline as the disease progresses. Pending the clinical manifestation and immune response, the symptom onset and infectious period can be greater than 14-days. For patients with mild-moderate illness, the duration of symptoms is typically 10-days [6].

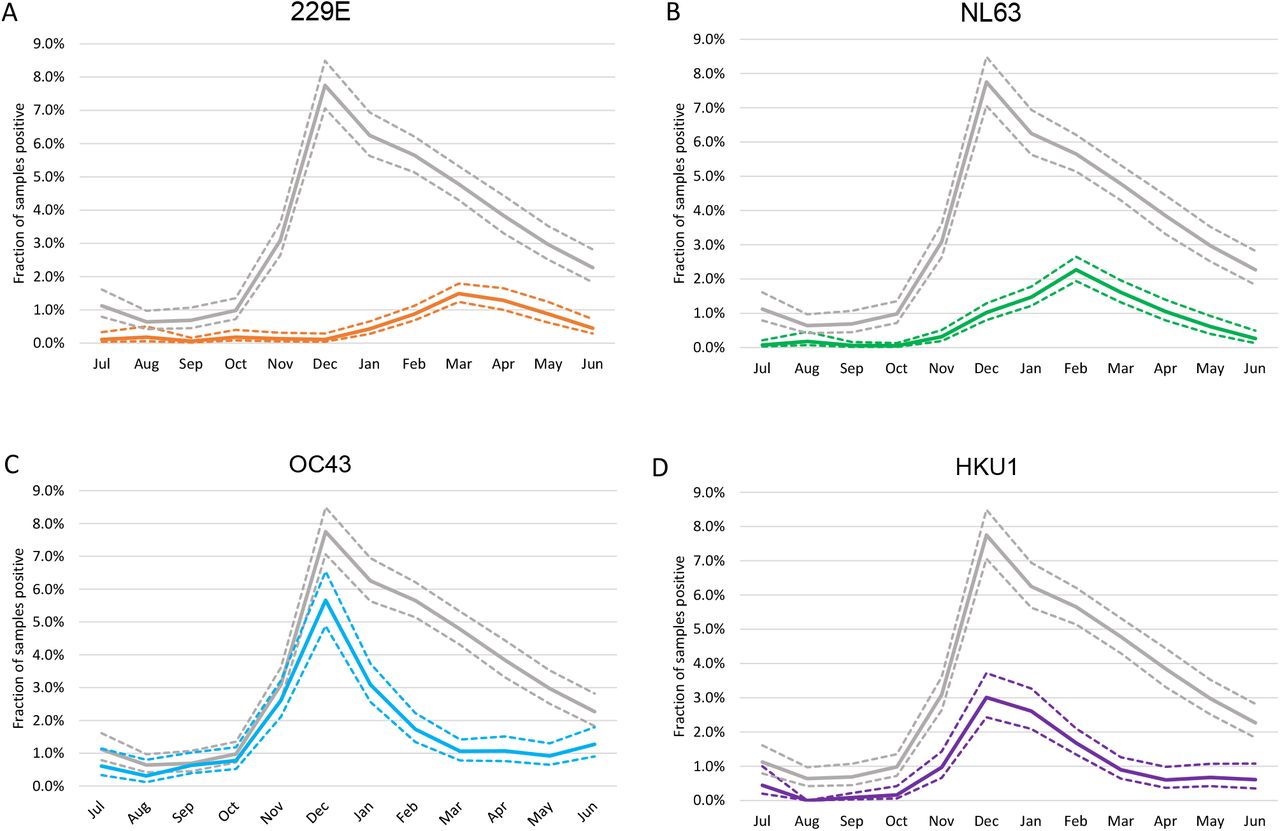
**Viral Load**

Viral load (VL) refers to the amount of virus detected through polymerase chain reaction (PCR) assay, which have been associated with transmission risk and disease severity [7]. Viral load can be detected through testing of various locations and materials including nasopharynx (nasopharyngeal swab), saliva, sputum, and stool. Through PCR analysis of the nasopharyngeal swab, the highest viral load is detected at the time of symptom onset, with a steady state decline [8]. The peak viral load in the upper respiratory tract occurs 2-4 days post symptom onset, and in severe illness, a viral load peak will occur in the lower respiratory tract around day 10- 11 [8].

**Seasonal Coronaviruses and PCR panels**

Various strands under the *Coronavirinae* family such as HCoV-229E, HCoV-NL63, HCoV-OC43, and HCoV-HKU1 were found to be seasonal [10]. Seasonal coronaviruses, 229E and NL63, were found to be the most present in Spring, while OC43 and HKU1 were most present in Winter, as shown in **Figure 1**. Testing for seasonal coronaviruses is completed by PCR respiratory panels with gene regions specific to the virus. These PCR panels are multiplexed and often screen for select infections common during a selected time of year (ie. Influenza, respiratory syncytial virus, adenovirus, parainfluenza, rhinovirus)[12]. The viral panel is dynamic; therefore, if a viral infection is not common during a specific season, testing will not be completed. Viral panels are a common practice in hospitals and clinical labs which help aid in infection epidemiology. As shown in **Figure 1**, the fraction of samples being tested, represented by the gray dotted lines, is consistent across all coronaviruses. This suggests that a PCR-based testing panel is consistent between the selected coronavirus, over the duration of various months.

To compare infection rates of seasonal coronaviruses, a respiratory panel including all seasonal coronaviruses would be needed. This would remove the testing selection bias and provide a comprehensive analysis of epidemiology between each seasonal infection, including SARS-CoV-2.



**Figure 1**: **Seasonal Peaks of HCoV rates including HCoV-229E, HCoV-NL63, HCoV-OC43, HCoV-HKU1**[11]. Figure A and B, 229E and NL63, respectively, will show increasing infection rates during the late winter and early spring. Where Figure C and D, OC43 and HKU1, respectively, will peak in infection rates during the late Fall and early Winter months. Dotted gray lines represent the fraction of samples being tested.

**Methods:**

*Data Collection:*

The samples were collected via nasal and/or nasopharyngeal swabs and were sent to the clinical lab at the Kingston Health Sciences Center for processing using qPCR. All samples with a CT of less than 40 were pooled for data analysis.

Along with the qPCR results, the SARS-CoV-2 dataset used for analysis was pooled with additional information of a Unique ID number, testing sample, biological age, sex, viral load, and collection date.

*Data Manipulation:*

All samples were grouped by Unique ID number which corresponds to a single SARS-CoV-2 positive individual. Samples with inconsistent SARS-CoV-2 positive data was cleaned for continuous result data (ie. gsub(pattern="positive", replacement = "Positive", df)). Samples were then removed with multiple Unique ID numbers leaving only one sample per individual. Sex stratification was completed by grouping ‘F’, ‘M’, and ‘NA’ for Biological sex analysis. Age cohort bins were analyzed which included <19, 19-64, and 65+. Finally, data was stratified into year of infection, by introducing a variable containing Year, using a group\_by function as shown:

Df <- Df1

mutate(Year = as.Date(day, format="%Y-%m-%d")) %>%

group\_by(Year) %>% # group by the day column

summarise(sum(COVID=="Positive"))

*Customizing Age Bin*

The age bin was then broken down into 10-year periods to check for trends across more specific ranges. A scatter plot was used to visualize the data, and a linear regression line was added to the plot after creating a linear model between age and the viral load. A scatter plot was more appropriate than a bar graph, because an equal number of data points were not available for each age bin. The correlation coefficient was also calculated for theoretical evidence.

*Seasonal impact on Viral Load*

To see how the viral load detected changed with the seasons, an appropriate data frame was first set up. This data frame had the rows with repeated UniqIDs and missing values removed. It also only included the rows where the test results were Positive. A new column was added to the dataset by creating a custom function to choose the season based on the summer and winter solstice. The ranges for each season are given below:

**Spring**: March 20 - June 20

**Summer**: June 21 - September 20

**Fall**: September 21 - December 20

**Winter**: December 21 - March 19

This new column was then used to plot a graph between the viral load and the season for the years 2020 and 2021 separately and together. ANOVA was carried out to check if there was a significant difference in viral load detected between the four seasons across the two years. Other variables such as biological age and sex were also added in to the intial plot to see if there was consistency in results across different demographics. The packages used for this analysis were dplyr, lubridate, and ggplot2.

*Generalized additive mixed model (GAMM)*

Generalized additive mixed modeling (GAMM) allows for trends to be plotted, with the ability for variances in the data to be taken into account. In this case GAMM was used to look at seasonality trends of time series data (COVID cases per day). The smoothing function used was lowess. A model checking function was created and used to determine the best “k” value for each dataset (2020,2021). A k value of 7 was chosen as it produced the best R squared value. Next we created a dataset of n=200 (based on the given COVID data) to look at periods of significant change (gam.d <- Deriv(gamm\_covid$gam, n = 200)). We set up a dataframe for the predicted values to be added to in the next line, and added the fitted GAMM values to the new dataset. The final figures were generated by plotting this new dataset along with the GAMM, and highlighting periods of significant increase in blue and periods of significant decrease in red. This analysis was done using the Hmisc and mcgv packages in R.

Some limitations of GAMM analyses include: violation of normality which can be mitigated by applying a model checking function and observing the residuals. Violation of homogeneity which can be addressed by using data transformation. Violation of independence can be addressed by incorporating a temporal or spatial dependence structure between the observations (or residuals) in the model. When taking on this type of analysis it is important to be aware of its limitations, and take precautions to mitigate them.

*Violin Plots:*

Another analysis similar to the dot plots is known as a violin plot: a hybrid of a box plot and a kernel density plot, which shows peaks in the data. It is used to visualize the distribution of numerical data. Violin plots were used to analyze demographic parameters (E.g., Gender and Age) within two different time frames: 2020 and 2021.

*Statistical Analysis:*

All data was statistically analyzed and coded using RStudio 9.3, R-script and Python 3.9. Coding analysis can be found on GitHib (Andrew add in repository here)

**Results:**

*Data Breakdown*

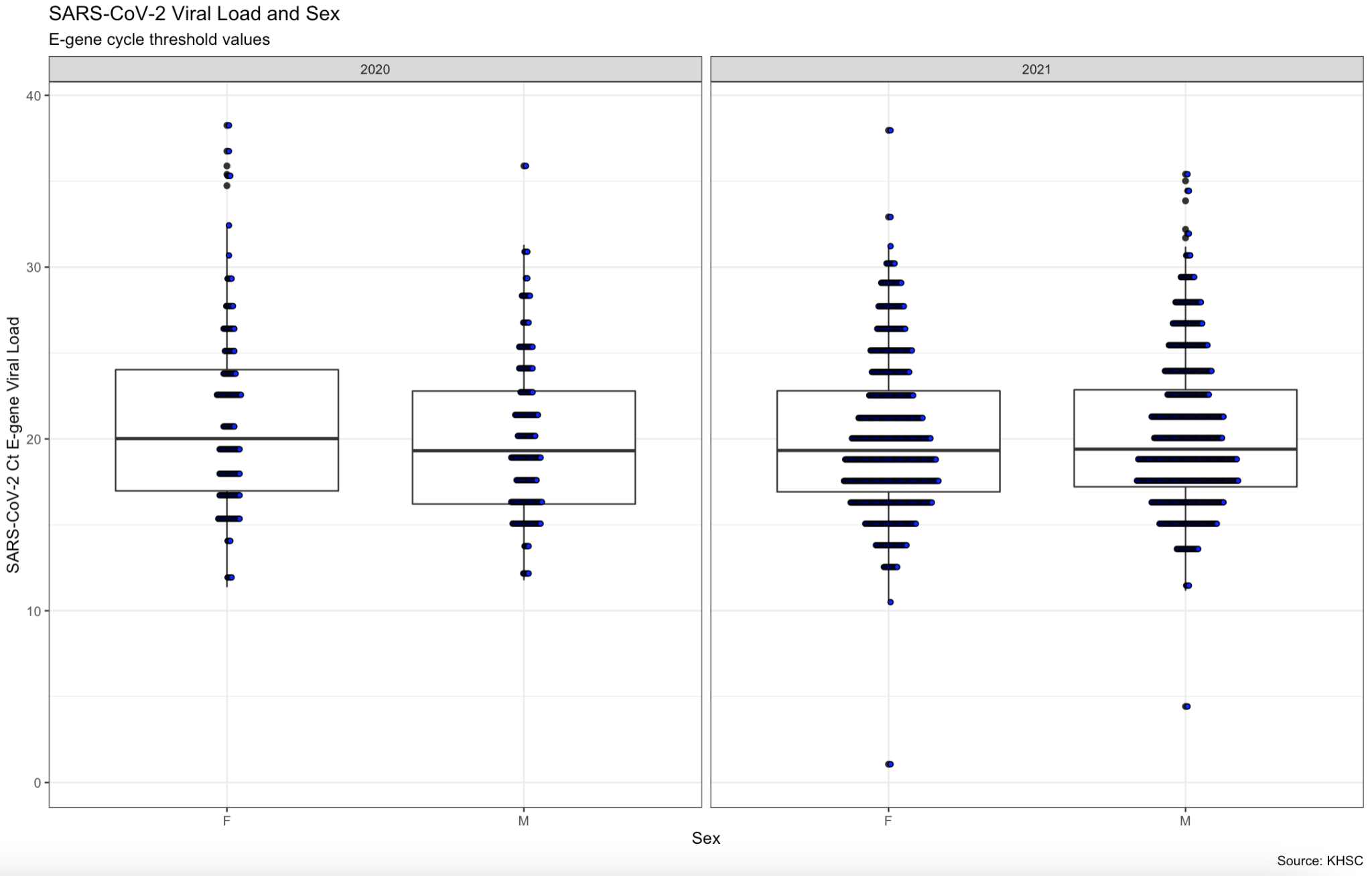
A total of 244.420 samples were used in this analysis. Specimens were assigned a unique ID number which corresponds with their selected sample. Individuals with multiple PCR samples, tagged with the same Unique ID, were filtered for a total number of n=2296 positive SARS-CoV-2 for this analysis. When stratifying based on Age, a maximum age of 99.50 years was observed, versus a minimum age of 0.533 (median 43.12). Biological sex shows X were male versus X female. Finally, when observing Cycle thresholds (Ct) as a surrogate marker for viral load, a maximum viral load was observed with 1.07 Ct’s and a minimum of 38.25 Ct’s (median 19.42), as shown in **Table 1.**

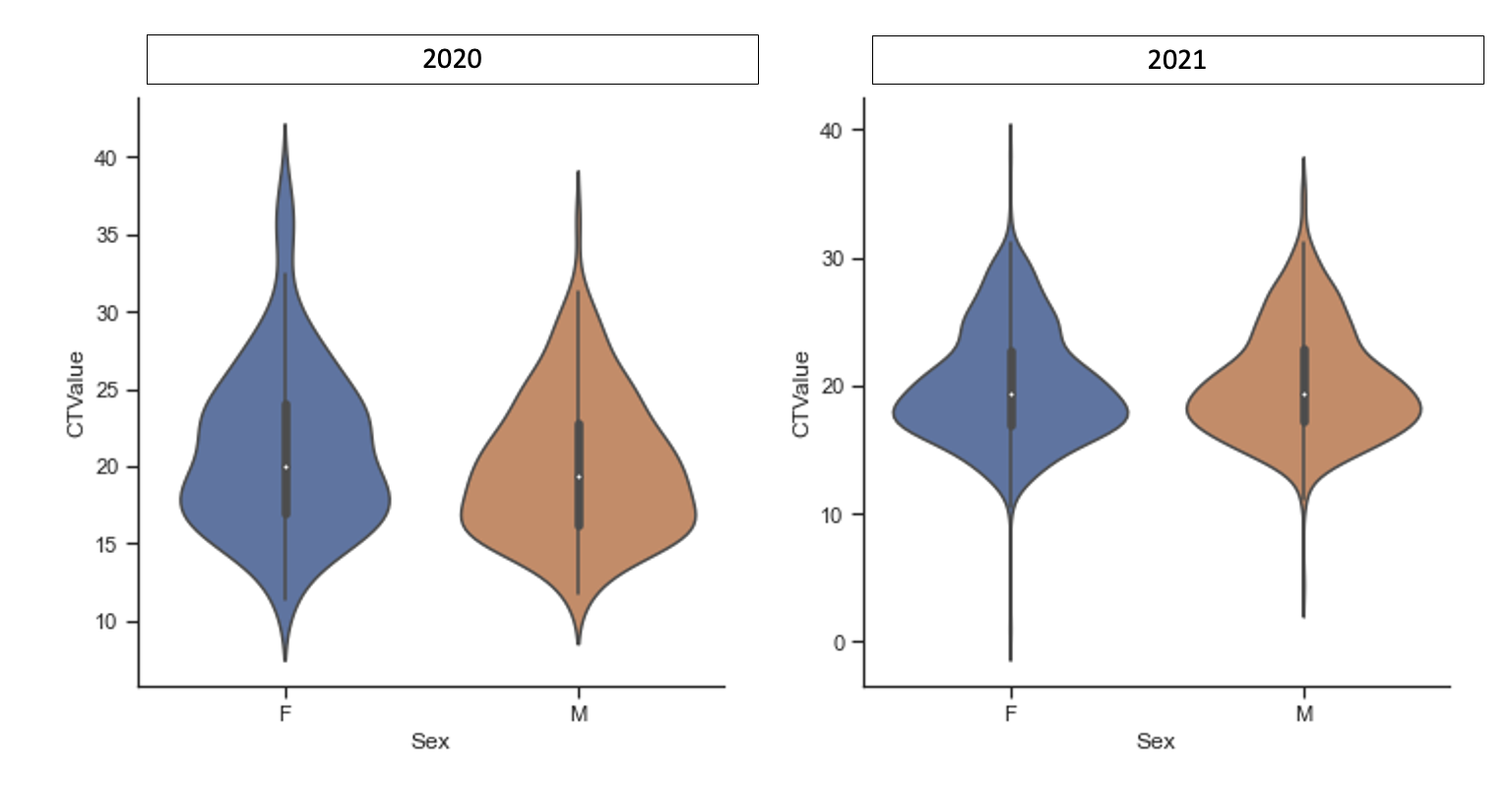
**Table 1**: **Demographic breakdown of dataset.** Values include Unique identifier (n=2296), COVID result, Age, Sex, and Ct viral load.

| **Variable** | ***n* =** | **Quartiles** | **Statistical Analysis completed** |
| --- | --- | --- | --- |
| Unique ID | 2296 | *Not applicable* | *Not applicable* |
| COVID result | 2296 | *Character* | GAMM |
| Age | 2140 | Max - 99.50  Min - 0.533  Median - 43.12 | ANOVA |
| Sex | 2215 | *Character* | Wilcoxon |
| CT Viral Load | 1446 | Max - 1.07  Min - 38.25  Median - 19.42 | ANOVA  Wilcoxon |

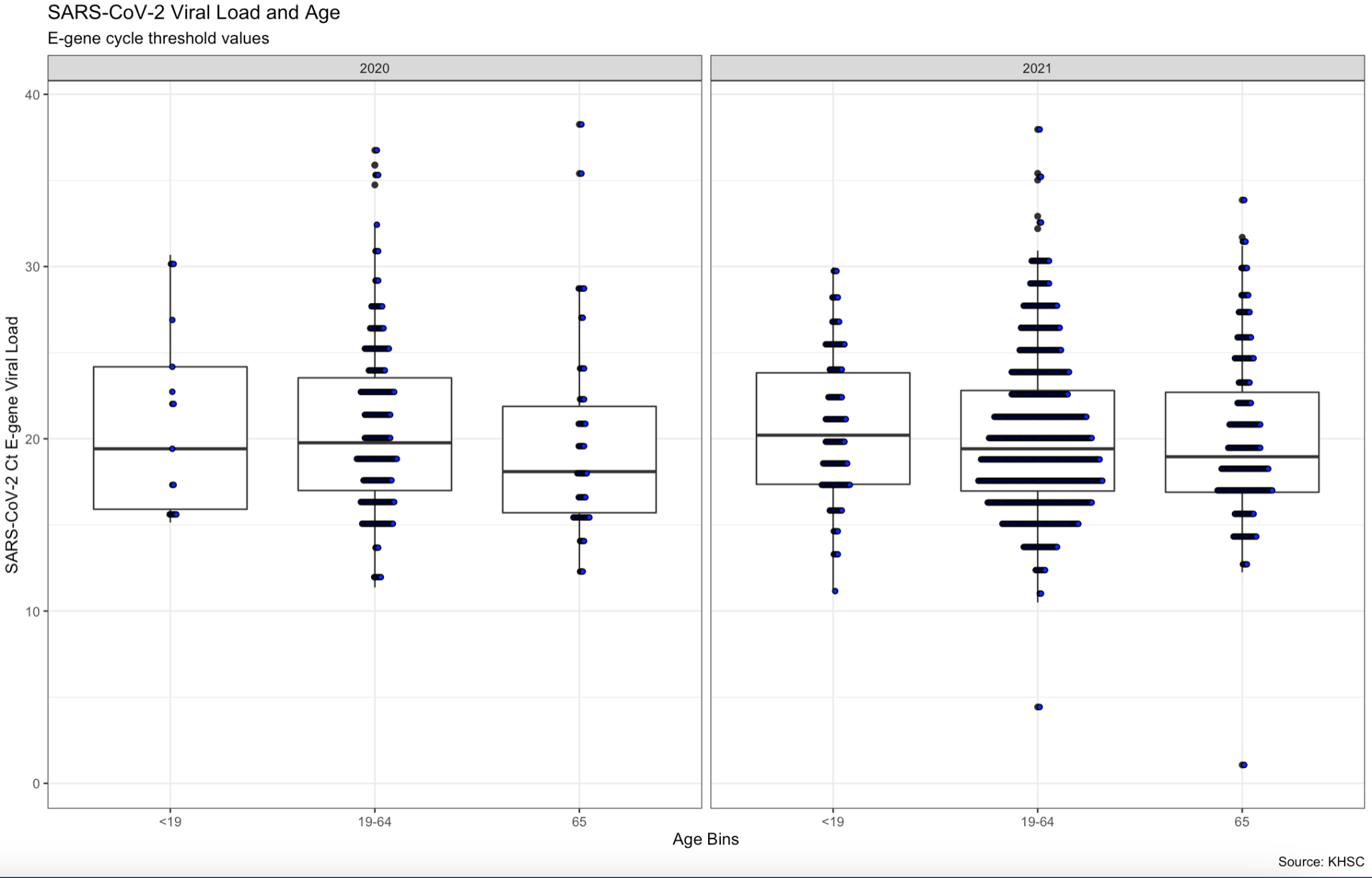
*Viral Load on Biological Sex and Age*

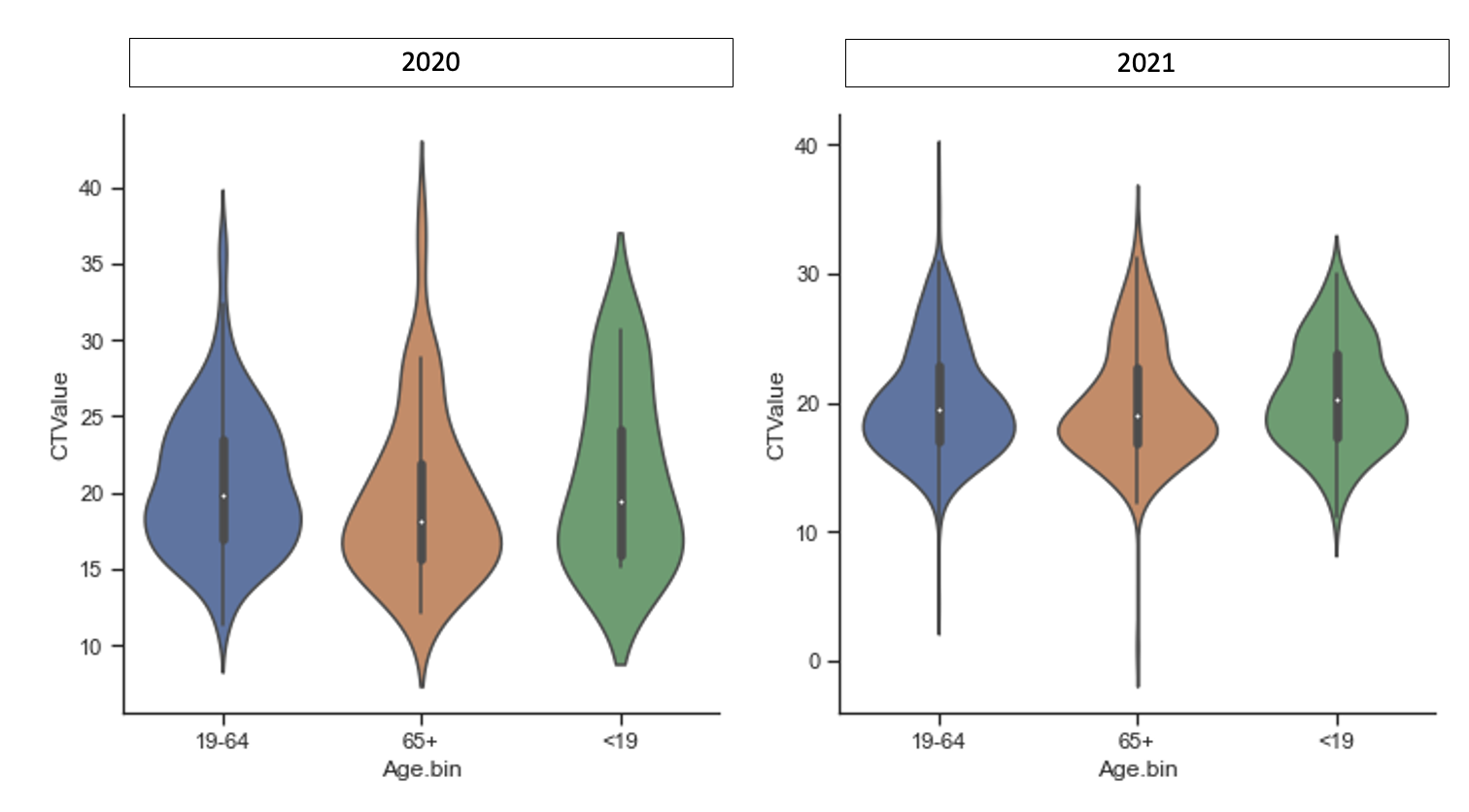
Using Cycle thresholds (Ct) values of the E-gene, of all positive SARS-CoV-2 individuals were analyzed. Biological sex and viral load were investigated using a Wilcoxin signed-rank test was used to compare the median threshold values. When stratifying based on 2020 and 2021 samples, we saw no statistical significance between males and females in both 2020 and 2021, as shown in **Figure 2A and 2B** (*p*=0.8711). An age cohort analysis of Ct E-gene was conducted and analyzed using an Anova model comparing the median rank of each column between 2020 and 2021. There was no statistical significance between Age and Viral Load in 2020 or 2021, as shown in **Figure 3A and 3B**. As Age Bins are often the primary way to represent age we suggest this to be a limitation as this does not provide a whole view of age and Ct viral load, rather binned groups and their median thresholds.

****Figure 2A**: **SARS-CoV-2 Viral Load E-gene Ct value and Biological Sex stratifyed by 2020 and 2021 via Dot Plot in R-Studio 9.3**. Using a Wilcoxon signed-rank test there was no statistical significance in median threshold values between females and males in both 2020 and 2021.

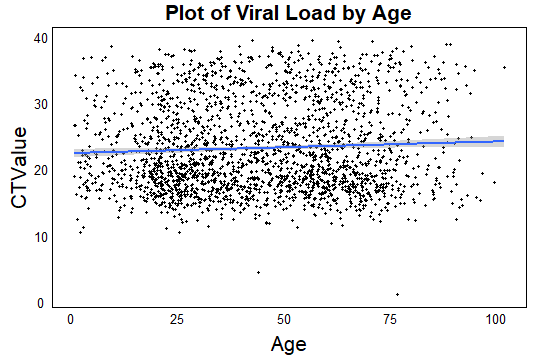
**

**Figure 2B**: **SARS-CoV-2 Viral Load E-gene Ct value and Biological Sex stratifyed by 2020 and 2021 via Violin Plot in Python 3.9.**

**

**Figure 3A**: **SARS-CoV-2 Viral Load E-gene Ct value and Age Cohorts stratified by 2020 and 2021 via Dot Plot in R-Studio 9.3.** Using an ANOVA model there was no statistical significance between Age Cohorts in both 2020 and 2021. 

**Figure 3B**: **SARS-CoV-2 Viral Load E-gene Ct value and Age Cohorts stratified by 2020 and 2021 via Violin Plot in Python 3.9.**

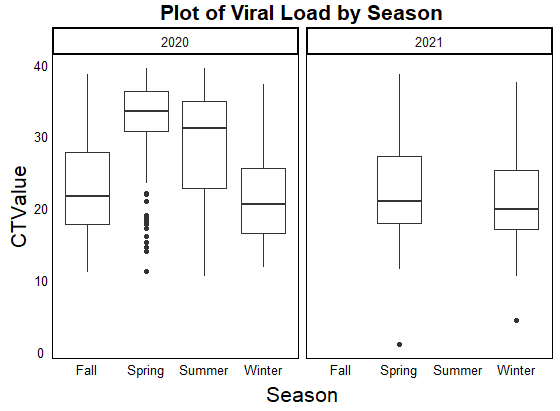
**

**Figure 4:** **Scatter Plot of Ct E-gene and Age**. Using a regression analysis there was no statistical significance of age on viral load, with a low correlation coefficient of 0.057.

The regression line as seen in **Figure 4** is almost parallel with the x-axis, indicating that Age has little impact on the viral load. The correlation coefficient was calculated to be 0.05668156 which is much less than 1, and is further proof of the result.

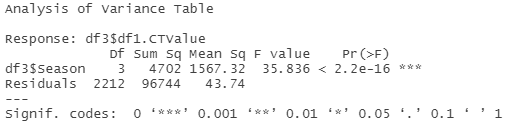
*Seasonal Impacts of Biological sex and Age on Viral Load*

As observed, in **Figure 5**, the viral load has an increase during spring and the summer of 2020. A similar trend is not seen in 2021. This may be due to different lineages of the virus causing the cases in the two years. The viral load is affected by a number of factors such as the sample type, the individual’s physiology, genetics, and collection date post infection. Since the seasonality of the SARS-CoV-2 virus has already been established[9], it might be intuitive to reason that a higher viral load is required for the virus to thrive in its off-season periods, i.e., Spring and Summer. It must be noted that only 1 data point is present in winter 2020, and data is not present for Summer and Fall 2021.

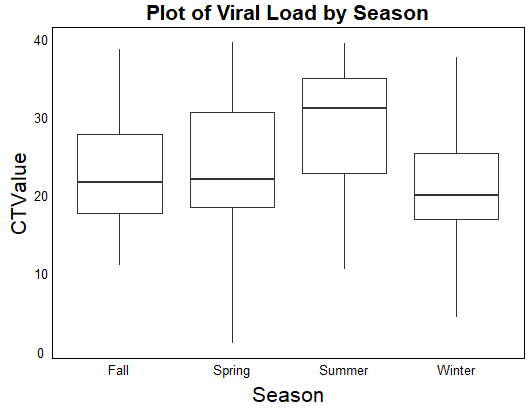


**Figure 5:** **Seasonal Impacts of Viral Load**. Using an ANOVA analysis of variance, there is a statistical significance in viral load between the seasons (*p*=<0.0001).

An ANOVA analysis of variance table, as shown in **Figure 6**, shows a statistical significance in viral load and season, indicating that there is some significant difference across the seasons. The ANOVA test might not have been the most appropriate for this dataset as it has uneven rows of data for the four seasons. The variance associated with each of the seasons were comparable, as seen in **Figure 7**, and acknowledging that even row size is not an assumption in ANOVA, the more relevant Kruskal-Wallis test was not carried out. This table was formed using data that was incomplete for the Summer and Fall of 2021, so there can be some skepticism around the result. However, it does give us some general idea about the seasonality associated with viral load, and is something that is definitely worth exploring further by taking more factors such as viral lineage into account.

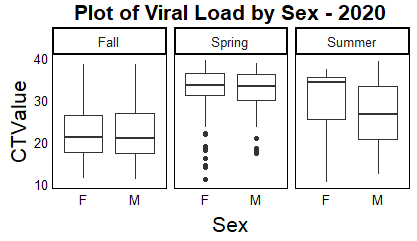


**Figure 6: ANOVA Analysis of Variance Table between Season and Viral Load.** Using an ANOVA there was a statistical significance between season and viral load (*p=* <0.0001).

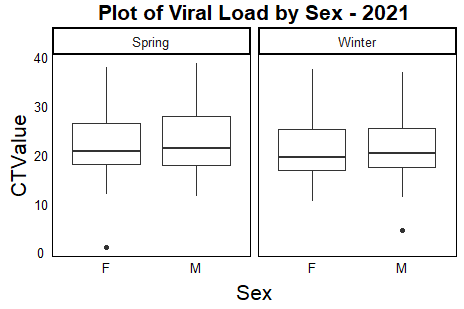


**Figure 7:** **Plot of Viral Load by Season including 2020 and 2021.** Identifies an increase in E-gene viral load in the Summer season with a decrease in the Winter months.

Further we explored the impacts of biological sex on viral load, as shown in **Figure 8**, across the seasons in 2020 and 2021, as seen in **Figure 9**. We see consistent trends of increase in VL during Spring and Summer for both the sexes. It is seen that females show higher VL in Summer of 2020 compared to males. It is unclear why this is the case; further correlation must be established after collecting data for Summer 2021.

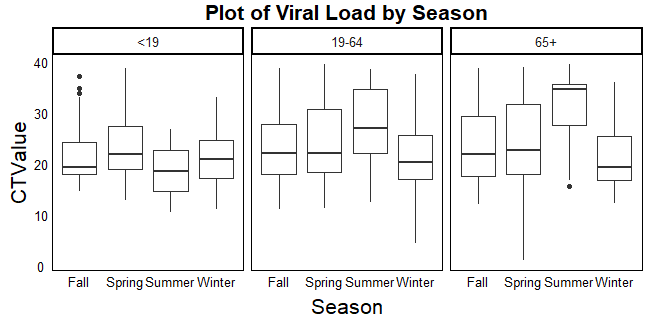
**

**Figure 8**: **Seasonal Impacts of Viral Load on Biological Sex in 2020.** Identifies an increase in E-gene cycle threshold values in Spring and Summer months for both males and females.

**

**Figure 9: Impacts of Biological sex on Viral Load in 2021.** Further identifies an increase in E-gene cycle threshold values in Spring and Summer months for both males and females.

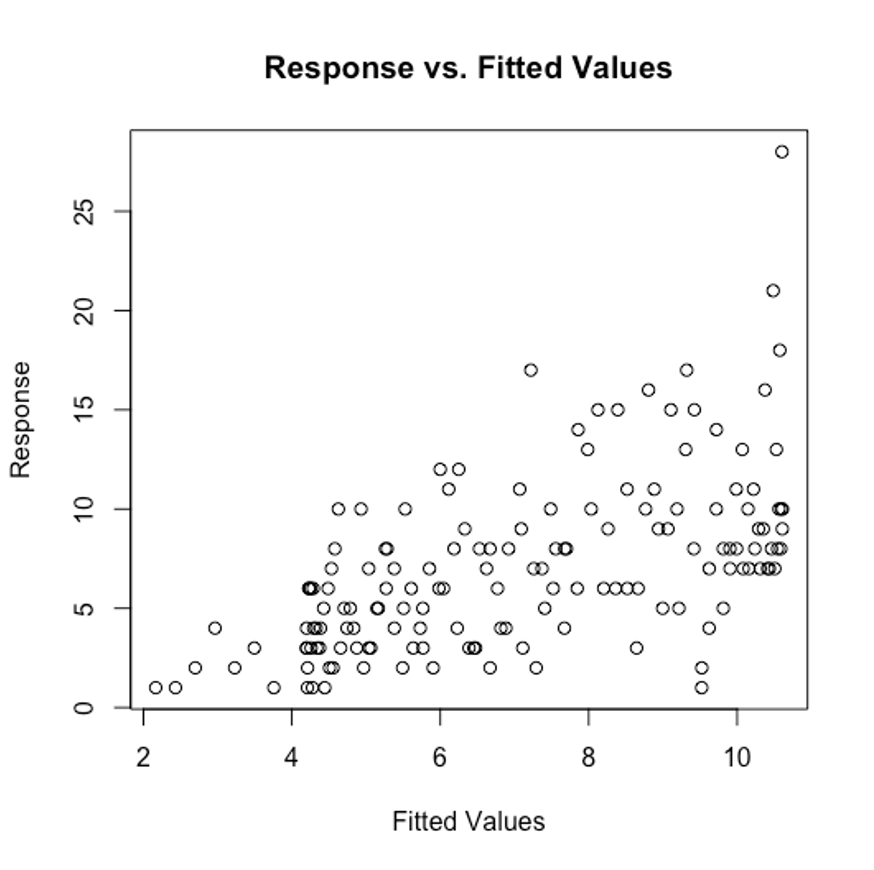
The viral load is compared between different Age cohorts across the four seasons in both the years, 2020 and 2021 in **Figure 10**. We see similar trends of increase during Spring/Summer across all age groups. There is an unusual decrease in VL during Summer in ages <19, and an unusual increase in VL during Summer in ages >65. The data points for these two age groups are not as complete as those available for ages 19-64. Unless there is some reasoning behind these results, it is quite possible that these anomalies might be errors due to incomplete data.

**

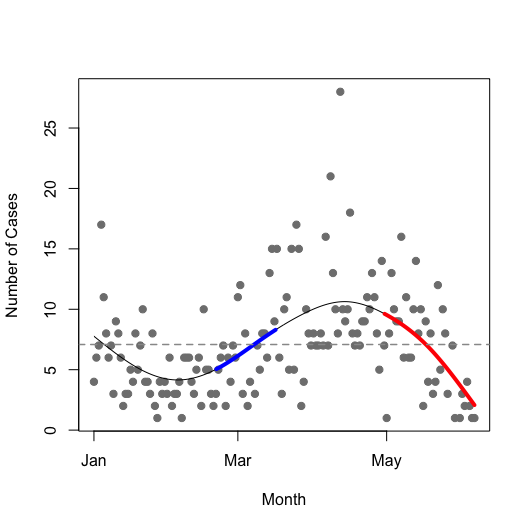
**Figure 10: Impacts of Age on Viral Load in 2020.** Identifies an increase in E-gene Ct values in Spring and Summer across all age groups. However, a decrease in viral load in the Summer months was observed in <19 where an increase was seen in 65+.

*Seasonal Impacts on Infection Rates*

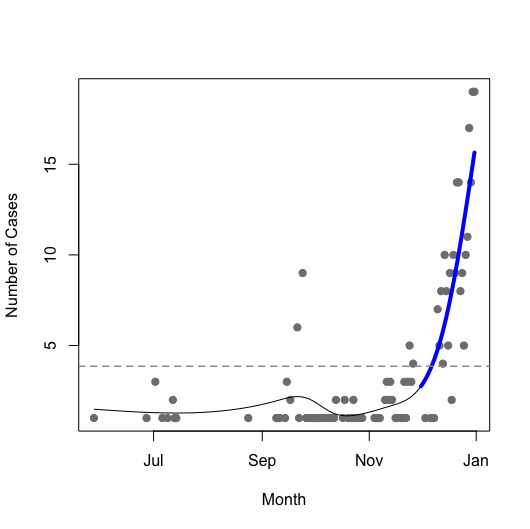
Seasonal changes are cyclic, predictable and influence almost all human and natural factors; many pathogens vary directly with seasonality (Altizer et al., 2006). Seasonal changes in host interactions is a large factor that contributes to the seasonality of pathogens. As well as variation in temperature, rainfall, humidity, seasonal timing of reproduction and pulses of susceptible hosts, seasonal changes in host susceptibility and immune defense (Altizer et al., 2006). One approach to characterize the seasonality of viruses is to incorporate seasonal forcing into models and/or mathematical models. We described the seasonality of our time series data using a generalized additive mixed model (GAMM). Seasonal shedding of viruses and seasonality has been shown to be highly synchronized, our GAMM analysis clearly shows that there is an increase in SARS-CoV-2 infection rates corresponding to the Spring months.

**

**Figure 11: Response vs Fitted Values using a Generalized Additive Mixed Model (GAMM).** Using a GAMM response for best fit, gam.check(gamm\_covid$gam) identifies k=7 as the best fit for modeling 2020 and 2021 SARS-CoV-2 infection rates (K=7).

**

**Figure 12: Trend Lines of SARS-CoV-2 Positive Infection Rates in 2020.** Using a GAMM model there is an increase in SARS-CoV-2 infection rates corresponding to the Spring months.

**

**Figure 13: Trendlines of SARS-CoV-2 Positive Individuals between June of 2020 and January of 2021.**

**Conclusion:**

E-gene Cycle thresholds (Ct) following the infection of SARS-CoV-2, appears to be associated with seasonal change. Our cohort had higher Ct values in the Spring and Summer months in both 2020 and 2021. We identified there was no statistical significance in Ct viral load between both biological sex and age. However, when stratifying based on seasonality we observed an increase in viral load in females in the Spring and Summer months. This was consistent across all age groups. This finding does not correlate with the current literature suggesting viral load impacts disease severity where the elderly population is at an increased risk of mortality and morbidity. Further, there is no literature showing an increase in viral load within the female cohort. These preliminary findings need further validation and suggest that host biological variables may play a role in SARS-CoV-2 infection rates. Finally, we suggest a quantitative analysis including a PCR respiratory panel targeting SARS-CoV-2, HCoV-229E, HCoV-NL63, HCoV-HKU1, and HCoV-OC43 throughout the duration of 12-months to further validate the infection rates between coronaviruses. This analysis will allow for a better understanding of the impacts of infection post pandemic. Lastly, this analysis will aid in public health measures when seasonal epidemics of SARS-CoV-2 occur.

**References**:

1. Altizer, S., Dobson, A., Hosseini, P., Hudson, P., Pascual, M., & Rohani, P. (2006). Seasonality and the dynamics of infectious diseases. *Ecology Letters*, *9*(4), 467–484. <https://doi.org/10.1111/j.1461-0248.2005.00879.x>
2. Wang, Huihui et al. “The genetic sequence, origin, and diagnosis of SARS-CoV-2.” *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology* vol. 39,9 (2020): 1629-1635. doi:10.1007/s10096-020-03899-4
3. Coronavirus Cases. Retrieved from, https://www.canada.ca/en/public-health/services/diseases/coronavirus-disease-covid-19.html
4. Fehr, Anthony R, and Stanley Perlman. “Coronaviruses: an overview of their replication and pathogenesis.” *Methods in molecular biology (Clifton, N.J.)* vol. 1282 (2015): 1-23. doi:10.1007/978-1-4939-2438-7\_1
5. Dorothea Bestle, Miriam Ruth Heidl, Hannah Limburg, Thuy Van Lam van, Oliver Pilram, Hong Moulton, David A. Stein, Kornelia Hardes, Markus Eickmann, Olga Dolnik, Cornlius Rohde, Stephan Becker, Hans-Dieter Klenk, Wolfgang Garten, Torsten Steinmetzer, Eva Bottcher-Friebertshauser. “TMPRSS2 and furin are both essential for proteolytic activation of SARS-CoV-2 in human airway epithelial cells and provide promising drug targets”. bioRxiv 2020.04.15.042085; doi: <http://doi.org/10.1101/2020.04.15.042085>
6. Jian Shang, Yushun Wan, Chuming Luo, Gang Ye, Qibin Geng, Ashley Auerbach, Fang Li. Cell Entry Mechanisms of SARS-CoV-2. Proceedings of the National Academy of Sciences. May 2020, 117 (21) 11727-11734; doi: 10.1073/pnas.2003138117
7. Civik Muge, Kuppalli Krutika, Kindrachuk Jason, Peiris Malik. Virology, transmission and pathogenesis of SARS-CoV-2 *BMJ* 2020; 371 :m3862
8. Elisabet Pujadas, Fayzan Chaudhry, Russell McBride, Felix Richer, Shan Zhao, Ania Wajnberg, et al. SARS-CoV-2 viral load predicts COVID-19 mortality. The Lancet Respiratory Medicine. August 06, 2020. Doi:<http://doi.org/10.1016.S2213-2600(20)30354-4>
9. Wölfel, R., Corman, V.M., Guggemos, W. *et al.* Virological assessment of hospitalized patients with COVID-2019. *Nature* **581,** 465–469 (2020). https://doi.org/10.1038/s41586-020-2196-x
10. Xiaoyue Liu, Jianping Huang, Changyu Li, Yingjie Zhao, Danfeng Wang, Zhongwei Huang, Kehu Yang, The role of seasonality in the spread of COVID-19 pandemic,

Environmental Research, Volume 195, 2021, 110874, ISSN 0013-9351, https://doi.org/10.1016/j.envres.2021.110874.

1. Interactions between seasonal human coronaviruses and implications for the SARS-CoV-2 pandemic: A retrospective study in Stockholm, Sweden, 2009-2020. Robert Dyrdak, Emma B. Hodcroft, Martina Wahlund, Richard A. Neher, Jan Albert. medRxiv 2020.10.01.20205096; doi: <http://doi.org/10.1101/202005096> now published in *Journal of Clinical Virology* doi:10.1016/j.jcv.2021.104754
2. Making sense of respiratory viral panel results. Retrieved from, https://asm.org/Articles/2020/March/Making-Sense-of-Respiratory-Viral-Panel-Results