Data Analysis Project

Leah Lariscy

3/18/23

# 1. Summary/Abstract

*Write a summary of your project.*

# 2. Introduction

## 2.1 General Background Information

Infection by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), causing COVID-19, is followed by the shedding of viral particles by multiple excretory functions, including stool and urine production. Subsequently, these viral particles can be detected in wastewater influent via RNA extraction, followed by Reverse Transcriptase Quantitative Polymerase Chain Reaction (RT-qPCR) utilizing SARS-CoV-2 specific primers. Wastewater-based detection methods have been utilized across the globe as an independent and parallel indicator of SARS-CoV-2 viral prevalence and, depending on community-specific factors, as a leading indicator of clinical case trends. WBE methods may also become a primary tool in understanding transmission dynamics as case under-reporting increases.

## 2.2 Description of data and data source

24-hour composite wastewater samples were collected twice weekly and refrigerated until ready for extraction. RNA extractions were performed for each sample collected, in replicates (n=6). Reverse Transcriptase Quantitative Polymerase Chain Reaction (RT-qPCR) and SARS-CoV-2 specific primer-probe pairs were used to amplify each RNA extraction replicate in technical replicates (n=3).

The concentration per reaction of each gene target was determined by the standard curve calculation of the positive control serial dilution corresponding to each target. Sample concentration in copies per liter was determined by Equation 1. Total daily copies per day per WRF was determined by Equation 2.

Insert equations here

## 2.3 Questions/Hypotheses to be addressed

Is there a relationship between viral load and clinical case observations?

Is there a relationship between qPCR percent positive and clinical case observations?

What do viral load trends look like over time?

What do clinical case trends look like over time?

How do each of these temporal trends compare to one another over time?

To cite other work (important everywhere, but likely happens first in introduction), make sure your references are in the bibtex file specified in the YAML header above (here dataanalysis\_template\_references.bib) and have the right bibtex key. Then you can include like this:

# 3. Methods

## 3.1 Data aquisition

Wastewater SARS-CoV-2 data was obtained through weekly surveillance efforts utilizing RT-qPCR methods.

Water Reclamation Facility data was obtained through data share directly from the WRF representatives.

COVID data downloaded from Georgia DPH website.

## 3.2 Data import and cleaning

### 3.2.1 RT-qPCR data

For each target, viral copies per microliter (cp/uL) of each PCR reaction was estimated by transforming the cycle threshold value (Ct) using the appropriate standard curve equation. From this, N1 and N2 data frames were combined and copies per liter (cp/L) of sample for both targets was estimated using Equation 1.

### 3.2.2 Treatment plant flow data

For each collection date, corresponding wastewater influent flow data, including total daily flow volume and total suspended solids rate, were combined with the qPCR data. Viral load (total daily viral copies) was then estimated by multiplying concentration estimates by daily flow volume (Equation 2).

### 3.2.3 COVID testing data

COVID-19 data sets were downloaded from the Georgia Department of Public Health website (accessed Jan 2023), of which a subset was selected which includes symptom onset, PCR test, and PCR positive test data for each date since the first case was detected. These were then combined into one data frame, and filtered to include only data pertaining to Clarke County. Then, this was further subset to include only data corresponding with the wastewater surveillance period.

## 3.3 Statistical analysis

Simple analyses: Wastewater and clinical data over time, linear regression model of wastewater vs clinical case data

Full analyses: Wastewater data as predictor for clinical case outcomes

Steps I took to build and tune model

# 4. Results

## 4.1 Exploratory/Descriptive analysis

**?(caption)**

|  |
| --- |
| Distribution of recorded seven-day moving average for each day |

|  |
| --- |
| Distribution of viral load per sample collection per target |

## 4.2 Basic statistical analysis

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

**?@fig-result** shows a scatterplot figure produced by one of the R scripts.

|  |
| --- |
| Clinical case seven-day moving average over time |

|  |
| --- |
| Wastewater viral load over time |

|  |
| --- |
| Linear regression of viral load vs case seven-day moving average |

## 4.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

Example **?@tbl-resulttable2** shows a summary of a linear model fit.

**?(caption)**

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (Leek & Peng, 2015) discusses types of analyses.

These papers (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, Shen, & Handel, 2020) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# 6. References

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>

McKay, B., Ebell, M., Billings, W. Z., Dale, A. P., Shen, Y., & Handel, A. (2020). Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery. *Open Forum Infectious Diseases*, *7*(11), ofaa494. <https://doi.org/10.1093/ofid/ofaa494>

McKay, B., Ebell, M., Dale, A. P., Shen, Y., & Handel, A. (2020). Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients. *Proceedings. Biological Sciences*, *287*(1927), 20200496. <https://doi.org/10.1098/rspb.2020.0496>