Wastewater Surveillance for SARS-CoV-2 Data Analysis Project

Leah Lariscy

5/5/23

# 1. Abstract

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 and used to estimate the abundance of COVID-19 in a community. 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. Many surveillance groups have already shown wastewater viral load to be predictive of population clinical cases, so this study will evaluate the predictive power of alternative wastewater metrics. This study also aims to address the utility of SARS-CoV-2 wastewater surveillance to predict viral prevalence within a population once clinical reporting of COVID-19 declines significantly.

# 2. Introduction

## 2.1 General Background Information

At the time of writing it has been three years since the COVID-19 Pandemic was declared, followed by a lockdown across the globe. To date, the World Health Organization reports over 760 million confirmed cases and 6.9 million deaths globally (*WHO Coronavirus (COVID-19) Dashboard*, n.d.). The causative agent, Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2), is a positive-sense single-stranded RNA zoonotic virus that infects cells via its spike proteins or “crowns” (V’kovski, Kratzel, Steiner, Stalder, & Thiel, 2021). When infected by this airborne pathogen, people report experiencing a wide range of mild to severe symptoms, including shortness of breath, cough, head and body aches, and loss of taste or smell. Older adults and others with underlying issues are at higher risk of experiencing more severe outcomes, including death (“COVID-19 and Your Health,” 2020). However, long-term effects of COVID-19 on the body are not well understood yet and it may be years before enough research has been done to well-characterize these impacts. The effects of this pandemic have been far-reaching, with influence practically in every aspect of society, indicating a need to keep a close eye on this infectious disease. However, as the pandemic progresses, clinical reporting continues to dwindle, leaving researchers needing a better means of collecting data on the spread of the virus.

Early on in the pandemic, wastewater-based epidemiology (WBE) was identified as a promising method for detection of SARS-CoV-2 in a population. This is due to the ability of SARS-CoV-2 particles to shed from the body during and after infection through excretory functions, namely stool production (Wu et al., 2020). Viral particles in wastewater are able to be detected via Reverse Transcriptase Quantitative Polymerase Chain Reaction (RT-qPCR) by amplification of universally-conserved gene targets on the SARS-CoV-2 genome (Ahmed et al., 2020). The use of universal targets allows for the detection of the virus regardless of what variant is present. As of writing, over 4,000 sites worldwide have contributed to wastewater monitoring of SARS-CoV-2 (*COVIDPoops19*, n.d.). Not only is WBE an effective way of detecting viral presence, but quantification of wastewater viral load has been shown to correlate strongly with COVID-19 clinical cases when reporting efforts were strong, indicating the potential for it to be used as an alternative disease monitoring method.

## 2.2 Description of data and data source

### SARS-CoV-2 wastewater surveillance data

Sample collection for wastewater surveillance began June 30, 2020 and occurred twice weekly (excluding major holidays) through December 21, 2022. Raw wastewater influent was collected from three Wastewater Treatment Plants in Athens-Clarke County and viral RNA was extracted. RT-qPCR was used to calculate the cycle threshold (Ct) value of each reaction, which is used to estimate SARS-CoV-2 copies per liter of wastewater in the data cleaning process (Section 3.2).

### Wastewater treatment plant data

Wastewater influent flow data, including millions of gallons per day (MGD) and total suspended solids (TSS), were collected for each corresponding sampling date.

### COVID-19 clinical data

COVID-19 clinical data were obtained from the Georgia Department of Public Health website (“COVID-19 Status Report,” n.d.) which includes COVID-19 reported clinical case data for the state of Georgia. These data include reported cases, demographics, comorbidities, deaths, and more. However, not all data is necessary for analysis and will be subset to include only symptom onset, positive case, and administered test data for Clarke County. (Section 3.2).

## 2.3 Hypotheses to be addressed

This study aims to conduct a parallel analysis between SARS-CoV-2 wastewater surveillance data and reported COVID-19 clinical case data in Athens-Clarke County, Georgia. Wastewater quantifications and reported case data are expected to have a positive linear relationship and it is hypothesized that certain wastewater-clinical relationships will be stronger than others, see Table 1 for variables of interest. It is also hypothesized that correlations will begin to decrease as the time series progresses due to diminishing clinical reporting.

# 3. Methods

## 3.1 Data acquisition

### Wastewater surveillance data

24-hour composite raw wastewater influent samples were collected twice-weekly from three treatment facilities in Clarke County and stored at 4°C until ready for extraction. 280 𝜇L of wastewater in replicates (n=6) was aliquoted and direct RNA extraction (Zymo Research, R2042) was performed on the same day as sample collection.Each extraction replicate was reverse transcribed and amplified using RT-qPCR and a SARS-CoV-2 specific primer-probe panel (IDT, 10006713), done in technical replicates (n=3). As part of the qPCR panel, two gene targets on the N-gene of the SARS-CoV-2 genome were used for amplification. The N1 and N2 gene targets were chosen due to their conservation across variants and act as universal detectors of the virus [@]. Cycle threshold (Ct) values were then calculated for each PCR reaction, which were then used to estimate SARS-CoV-2 viral load (See Section 3.2 for more detail).

### Wastewater treatment plant data

For each treatment plant and collection date, MGD and TSS was shared by plant operators directly with our team and compiled throughout the study period.

### COVID-19 clinical data

Reported case and testing data were downloaded from Georgia Department of Public Health’s COVID-19 Status Report Page in January 2023.

## 3.2 Data import and cleaning

### Calculating wastewater viral load

Amplification standard curve equations where generated from serially diluted synthetic controls (Supplementary Table). This was done for N1 and N2 gene targets for both surveillance years because RT-qPCR methods were altered after the first year. Viral copies per microliter (cp/uL) of each PCR reaction was estimated by transforming the Ct value using the appropriate standard curve equation. For reactions with no detection, concentration was set to the theoretical limit of detection (0.004 cp/uL). Data sets for each standard curve was transformed separately before being combined. After combining all qPCR data, copies per liter (cp/L) of wastewater was calculated for each technical replicate (Equation 1). Average cp/L of wastewater was determined by taking the mean of all technical replicates per sample per gene target.

*Equation 1. Viral copies per liter of wastewater (C\_W) was calculated by first multiplying copies per qPCR reaction (C\_R) by volume of reaction (V\_Q) divided by volume of converted DNA (V\_D) in the reaction, then multiplying by volume of the reverse-transcription reaction (V\_R) divided by volume of RNA extract (V\_R) in that reaction, then multiplying by the elution volume of the RNA extraction (V\_E) divided by the total volume of wastewater input (V\_W). This yields copies per microliter of raw wastewater, which is finally converted to copies per liter.*

WWTP data was then transformed by converting the flow rate from millions of gallons per day to liters per day (Equation 2). qPCR data was then combined with WWTP data and total viral load per sample was calculated by multiplying average copies per liter by the flow rate. Average viral load for each sample was then determined by taking the mean of the viral load for both N1 and N2 gene targets. County-level daily viral load was calculated by taking the sum of the average viral load from each of the three WWTP samples.

*Equation 2. Wastewater viral load (VL\_W) was determined by multiplying viral copies per liter of wastewater (C\_W) by the total daily volume of influent flow in liters (V\_I).*

### Calculating wastewater qPCR assay positivity

qPCR assay positivity was determined for each collection date by finding the proportion of assays that detected SARS-CoV-2 out of the total number of assays performed. This includes all biological and technical replicate assays that were performed for a given sampling date. County-level assay positivity data was then combined with county-level viral load data.

### Adding COVID-19 clinical data

Of the downloaded Georgia DPH data, the data set containing positive case counts was retained. Of these data, daily reported case counts and case seven-day moving average were kept for analysis. These data were then subset to include only data from Clarke County collected during the surveillance period and combined with wastewater data.

## 3.3 Statistical analysis

Through exploration of time series and wastewater-clinical correlations, prominent relationships were identified for further analysis. This also allowed for identification of an appropriate data split, whereby earlier observations were used to predict later ones. Wastewater variables were then used to train single-variate linear regression models to predict clinical observations. Multi-variate wastewater models were also created in the same manor to predict clinical observations. Model performance metrics (RSQ and RMSE) were assessed against each other and against a null model. Cross-validation was used to validate performances. The best model was then determined and used to predict values in data that was not used for training.

# 4. Results

## 4.1 Time-series analysis

Comparison of wastewater and clinical variable time-series showed that these variables, although independent of one another, follow the same trends over time (Figures 1 & 2, Supplementary Figures). However, this changes in the time-series of administered clinical tests after the first few months of 2022 where there is nearly a complete drop-off in testing (Figure 3). This observation led to the decision to train linear regression models on data collected prior to February 28, 2022 and predict data collected after this date.

|  | Overall |
| --- | --- |
|  | (N=224) |
| Log10 Viral Load |  |
| Mean (SD) | 13.7 (0.807) |
| Median [Min, Max] | 13.9 [12.1, 15.2] |
| Assay Positivity |  |
| Mean (SD) | 43.8 (32.3) |
| Median [Min, Max] | 39.3 [0, 99.1] |
| Clinical Case Seven-Day Moving Average |  |
| Mean (SD) | 11.7 (9.13) |
| Median [Min, Max] | 9.35 [0.200, 34.7] |
| Daily Reported Clinical Cases |  |
| Mean (SD) | 44.5 (62.4) |
| Median [Min, Max] | 25.0 [0, 365] |
| PCR Test Positivity |  |
| Mean (SD) | 10.3 (9.16) |
| Median [Min, Max] | 7.16 [0, 40.6] |

Table 1. Statistics on clinical and wastewater variables of interest

|  |
| --- |
| Figure 1. Wastewater qPCR assay percent positivity time-series |

|  |
| --- |
| Figure 2. Clinical case 7-day moving average time-series |

|  |
| --- |
| Figure 3. Daily COVID-19 tests administered time-series |

## 4.2 Correlation analysis

Two wastewater and two clinical case metrics were identified as features of interest during exploratory analysis steps where stronger wastewater-clinical correlations were observed between these variables than in others. These include wastewater viral load, wastewater assay positivity rate, clinical case seven-day moving average, and clinical test administration positivity rate (Table 1). Of these correlations, the strongest were observed when wastewater assay positivity was included as opposed to viral load (Figures 4 & 5, Supplementary Figures). Although this is true, linear regression models were still trained using both viral load and assay positivity.

|  |
| --- |
| Figure 4. Wastewater qPCR assay positivity vs clinical case 7-day moving average |

|  |
| --- |
| Figure 5. Wastewater qPCR assay positivity vs clinical test positivity |

## 4.3 Model analysis

Out of each feature combination, the model predicting seven-day moving average (7DMA) by wastewater assay percent positivity (APP) had the best performance metrics, confirmed by five-fold cross validation repeated five times (Table 2, Supplementary Tables). Metrics were also stable across a set of test data prior to February 28, 2022 that was withheld from model training. Thus, the 7DMA-APP model was selected to be fit to the test dates (after February 28, 2022). The metrics of this fit were also assessed, of which RMSE did not out-perform the null model (Table 3, Supplementary Table).

| Metric | Estimator | Mean | n | Standard Error |
| --- | --- | --- | --- | --- |
| rmse | standard | 6.442260 | 25 | 0.1611291 |
| rsq | standard | 0.464196 | 25 | 0.0218235 |

Table 2. 7DMA-APP Performance metrics when fit to training data

| Metric | Estimator | Estimate |
| --- | --- | --- |
| rmse | standard | 8.1181952 |
| rsq | standard | 0.3874664 |

Table 3. 7DMA-APP Performance metrics when fit to dates after February 28, 2022

|  |
| --- |
| Figure 6. Observed Clinical Cases vs Wastewater Predictions |

In Figure 6, 7DMA predicted by the APP model is shown in blue, whereas the reported value is shown in black. The closest predictions appear to take place towards the beginning of the time-series, whereas the gap in predicted vs actual value widens as time progresses. It also appears that in many instances of high case abundance, the predicted values are under-estimated. This is likely because at a certain point, the qPCR assays reach nearly 100% and cannot accurately estimate values above that threshold. However, it is interesting that in the Fall of 2023 when clinical cases appeared to be severely under-reported (Figure 3), cases are estimated to be nearly as high as during the prior peak during the late summer months, whereas case abundance in the reported data is seemingly much lower.

# 5. Discussion

## 5.1 Summary and Interpretation

There were more significant correlations determined between certain variables than others, but SARS-CoV-2 wastewater surveillance metrics overall were shown to be predictive of clinical trends. These findings also suggest that qPCR assay positivity could be an alternative measure to viral load in relating wastewater data to population transmission levels of COVID-19. This is valuable because this metric involves simpler calculations and could be implemented in instances where treatment plant flow data is unavailable. It is also worth noting that wastewater-clinical correlations were higher in the beginning of the pandemic when testing efforts were more robust, whereas correlations were weaker in later dates. This, coupled with the finding that case reporting in Clarke County significantly decreased after February 2022, implicates this method of wastewater surveillance as a reliable means of understanding community disease risk when clinical data is not.

## 5.2 Strengths and Limitations

The strength of this study lies in the ability of multiple wastewater metrics to predict clinical case data and that an alternative metric to viral load was identified. However, if improvements were to be made, more model tuning would be appropriate, seeing as certain model performance metrics did not far-exceed the null model’s performance. It is also worth noting that this could be partially due to the maxing out of assay percent positivity, making it difficult for the model to accurately estimate case abundance once the assay has become completely saturated. Because of this, this particular model may be more useful at predicting periods of high vs low case abundance rather than trying to accurately estimate case counts. If done again, I would categorize each date in the wastewater time-series as either high, medium, or low viral abundance based on the distribution of the data, and build a categorical prediction model to predict periods of high, medium, or low case abundance.

## 5.3 Conclusions

Overall, the content of this work highlights the efficacy of these wastewater surveillance methods in capturing accurate trends in clinical cases within this county. qPCR assay positivity was also determined to be a slightly better predictor of clinical data than wastewater viral load. Subsequently, a linear regression model was trained on qPCR assay positivity to predict clinical seven-day moving average and was applied to a time-series of low case reporting. Model performance metrics were better when the model was fit to the training data (prior to February 2022), but performance worsened significantly when fit to testing data (after February 2022). Due to upper-bound limitations though, the highest case levels were under-predicted both in the training and testing data. However, it is still worth noting that during the Fall of 2022 when case reporting was low but assay positivity was high, the model estimated that case levels were much higher than what was reported.

# 6. References

Ahmed, W., Angel, N., Edson, J., Bibby, K., Bivins, A., O’Brien, J. W., … Mueller, J. F. (2020). First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: A proof of concept for the wastewater surveillance of COVID-19 in the community. *Science of The Total Environment*, *728*, 138764. <https://doi.org/10.1016/j.scitotenv.2020.138764>

COVID-19 and Your Health. (2020). https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html.

COVID-19 Status Report. (n.d.). https://dph.georgia.gov/covid-19-status-report.

*COVIDPoops19*. (n.d.). https://www.arcgis.com/apps/dashboards/c778145ea5bb4daeb58d31afee389082.

V’kovski, P., Kratzel, A., Steiner, S., Stalder, H., & Thiel, V. (2021). Coronavirus biology and replication: Implications for SARS-CoV-2. *Nature Reviews Microbiology*, *19*(3), 155–170. <https://doi.org/10.1038/s41579-020-00468-6>

*WHO Coronavirus (COVID-19) Dashboard*. (n.d.). https://covid19.who.int.

Wu, Y., Guo, C., Tang, L., Hong, Z., Zhou, J., Dong, X., … Huang, X. (2020). Prolonged presence of SARS-CoV-2 viral RNA in faecal samples. *The Lancet Gastroenterology & Hepatology*, *5*(5), 434–435. <https://doi.org/10.1016/S2468-1253(20)30083-2>