Wastewater-Based Epidemiological Tracking of SARS-CoV-2: Determining Unreported Cases and Clinical Data Lag Time

Table of Contents

Abstract

Wastewater-based epidemiology (WWBE) is currently being studied as a promising future early warning and tracking tool for infectious diseases. WWBE uses viral RNA in sewage to track incidence and distribution of disease. This research asked two primary questions to evaluate the efficacy of WWBE: what is the approximate number of unreported cases in the Somerton municipality, and what is the approximate lag time between wastewater and clinical data? It was hypothesized that clinical cases represented less than a quarter of actual disease prevalence based on CDC data, and that lag time was within 6-14 days based on the current body of research. RT-QPCR was performed on wastewater samples to quantify SARS-CoV-2 RNA. For one sample per month, RT-QPCR was performed again to quantify pepper mild mottle virus for normalization. Resultant data was used to calculate estimated disease prevalence (EDP), the number of people at peak shedding rate. Monthly mean viral copies were also calculated to reduce noise. EDP case numbers, as hypothesized, were significantly greater than clinical case numbers. Normalized values did not follow expected results. They were much lower than non-normalized viral copies and appeared to have much more variation. Trends in normalized viral copies seemed to lag behind clinical trends, conflicting with the significantbody of research finding the opposite result. However, that WWBE detects COVID-19 cases before clinical data was strongly supported by more reliable data-- the comparison of monthly mean viral copies to clinical cases. Three peaks were defined in which lag between data sets was clear. Lag time can be approximated as one to three months. To address the unreported cases hypothesis, each clinical case value was compared to EDP. 88% of data gathered supported the hypothesis that clinical case numbers are less than a quarter of actual disease presence. Limitations of this research include the many unknowns/outside variables that are inevitable in field epidemiology, along with availability of clinical data. Future research inquiries may address fecal shedding differences by demographic and optimal normalization methods.

Introduction

Tracking the progression of the COVID-19 pandemic has been challenging since the beginning of its spread. The high number of asymptomatic and otherwise unreported cases have meant that researchers only have an approximate idea of the full scope of the pandemic. The CDC estimates that only 1 in 4 cases are reported, and that even that number is an underestimate of actual cases. For that reason, it is difficult to predict future peaks and troughs in disease prevalence and take public health actions based on clinical data.

Wastewater-based epidemiology (WWBE) is currently being studied as a promising future early warning and tracking tool for infectious diseases. WWBE uses viral RNA found in sewage to track the incidence and distribution of disease. It can be used to model occurrences and prevalence of disease in advance of clinical data. From the beginning of the COVID-19 pandemic, researchers began testing wastewater-based tracking tools with the novel coronavirus. One study (Medema et al) was able to predict the first positive test in the Netherlands six days prior to clinical data collection.

Today, polymerase chain reaction (PCR) testing is being performed on wastewater samples in cities across the globe to collect data on SARS-CoV-2 in viral shedding. Yuma is one of them--current sampling covers about 70% of the city population and the majority of the population in the surrounding municipalities (B. Schmitz & S. Slinski, personal communication, Oct. 18, 2021).

Ideally, based on fecal shedding rate, researchers could determine exact case numbers through PCR testing on untreated wastewater. However, determining the exact number of cases based on viral RNA fecal shedding is challenging, because many correlating factors could influence the viral RNA shed per patient. Individuals may shed different quantities of RNA based on case severity, demographic, progression, and other factors. Therefore, it may not be possible to determine an exact number of cases. Case numbers can, however, be estimated via methods based on research such as Schmitz et al, which quantified fecal shedding of viral RNA per person from wastewater samples of University of Arizona students.

The primary aim of this research was to estimate the number of unreported SARS-CoV-2 cases in a well-defined Yuma municipality (Somerton) based on viral RNA quantities in wastewater samples compared to clinical data. The secondary aim of this research was to determine the lag time between wastewater-based case numbers and clinical case numbers-- i.e., determining the effectiveness of WWBE as an early warning system. This can be accomplished by creating an epidemiological map and comparing it to clinical data. Somerton is the optimal municipality for this research project because it is a very well defined community that correlates easily with clinical data. The exact demographics, population, and clinical case numbers are known. All of the Yuma municipalities tend to behave similarly, so this research will focus on Somerton as a model system that can be generalized to the rest of Yuma.

Normalization of the SARS-CoV-2 RNA data will be performed to reduce noise and increase precision by estimating viral particles per person. Normalization scales data so that different arrays may be compared; it is a type of ratio metric comparison. PMMoV normalization has

proved effective for normalization of SARS-CoV-2 data (D'Aoust et al). PMMoV, or pepper mild mottle virus, is an RNA virus that affects plants and has no seasonality. It is abundant in human feces, being heavy in our diet, and passes straight through the digestive system. Based on methods refined in Zhang et al. and Haramoto et al., quantification of PMMoV in wastewater was used to estimate the number of individuals represented in each sample and thus scale the SARS-CoV-2 data to create more accurate comparisons between arrays.

Formulating a hypothesis for the primary question-- i.e., estimating unreported case numbers-- was difficult, given that there is not yet a reliable method to approximate unreported case numbers. The CDC has an estimate, but its accuracy is questionable, hence extensive research into WWBE. As stated, the CDC approximates that 1 in 4 cases are reported as of today, but claims that this estimate is likely low. Therefore, the hypothesis for the number of unreported cases in the Somerton area was some quantity greater than four times clinical case numbers at any given time.

The hypothesis for the secondary aim-- i.e., estimating lag time between clinical and WWBE case numbers-- was less challenging to formulate given that there is considerable applicable prior research available. Lag time between WWBE and clinical numbers is essentially the difference between when a person begins fecal shedding and when they become symptomatic to a degree significant enough that the individual gets tested. Based on the body of research available, lag time was hypothesized as somewhere within the range of 6-14 days (Peccia et al., Chavarria et al., Ahmed et al.). However, it must be noted that lag time does have an inherent error of a day or

two depending on excretion frequency and timing, as well as the time it takes for wastewater to get to the collection site/wastewater plant.

Methods

Effluent samples were collected from the single site in the Somerton municipality of Yuma County and transported on ice to the Yuma Center for Excellence for Desert Agriculture for experimentation, data collection, and storage. All data collection involving wastewater samples was conducted within the laboratory with proper protective garments, equipment, and workplaces. All student-conducted work was supervised by laboratory personnel trained in the methodology.

Each sample was subjected to stepwise filtration and centrifugal ultrafiltration to filter it down to approximately 200ul. Millicup-FLEX nonsterile, vacuum-driven filtration units (MilliporeSigma, Burlington, MA) were used for filtering directly into GL45 vacuum-rated filtration bottles. 47mm AAWP04700 pore size 0.8 um membranes were utilized, as well as sterilized support collars compatible with 47mm membranes. A vacuum pump/vacuum source capable of maintaining 25 in. Hg equipped with a shutoff valve and vacuum gauge with PVC tubing was attached to the Millicup. A 250 ml sterilized centrifuge bottle was also utilized.

After the filtration system was assembled, 0.8 um mixed cellulose ester membrane was placed on the support collar using sterile tweezers. The vacuum pump was then attached to the filter using PVC tubing, the pump was turned on, and the sample was slowly poured into the filter. If the filter slowed, it often indicated that it needed to be replaced. For highly turbid samples, multiple membranes were used. In this scenario, the membrane filter was aseptically removed using a sterile tweezer and replaced with a new membrane.

A Centricon-Plus 70 ultrafilter (MilliporeSigma, Burlington, MA) was utilized for the centrifugation-driven concentration process. The Centricon-Plus was pre-wetted by adding 50ml of Nanopure water, followed by centrifugation (1900 X g for 8 min). The unit was then inverted and centrifuged (800 X g for 2 min) to collect the remaining water, which was then discarded. Next, 70mL of the filtrate volume was transferred to the Centricon filter and concentrated via centrifugation (3500 X g for 30 min). In the final part of the concentration process, the viral concentrate was collected via inversion from the filter and further centrifuged for 5 min at 1000 X g.

200ul of the sample was purified. This 200ul was purified using the QIAamp Viral RNA Mini Kit (QIAGEN, Germantown, MD) according to manufacturer's directions.

SARS-CoV-2

Two sets of primers and probes were used for the SARS-CoV-2 qPCR, both obtained from the 2019-nCoV RUO Kit (Integrated DNA Technologies, Coralville, IA). These are as follows:

- 2019 nCoV_N1 Forward Primer
    - 5' - GAC CCC AAA ATC AGC GAA AT - 3'
- 2019 nCoV_N1 Reverse Primer
    - 5' - TCT GGT TAC TGC CAG TTG AAT CTG - 3'

- 2019 nCoV_N1 Probe
    - 5’ - FAM-ACC CCG CAT TAC GTT TGG TGG ACC-BHQ1 - 3’
- 2019-nCoV_N2 Forward Primer
    - 5’ - TTA CAA ACA TTG GCC GCA AA - 3’
- 2019-nCoV_N2 Reverse Primer
    - 5’ - GCG CGA CAT TCC GAA GAA - 3’
- 2019 nCoV_N2 Probe
    - 5’ - FAM-ACA ATT TGC CCC CAG CGC TTC AG-BHQ1 - 3’

The qPCR plate had 20ul of solution total per well. 14ul of Reliance One-Step Multiplex Supermix was pipetted into each well. In the wastewater wells, 6ul of sample was added per well, replicated 3x horizontally. In the positive control (PC) wells, 6ul of positive control solution (229E) was added per well, replicated 2x vertically. The 229E GBlock utilized for the positive control solution is as follows:

5’ - GGG CTA ATT GGG ACT CTA ATT GGG CCT TTG TTG CAT TTA GCT TCC TTA TGG CCG TAT CAA CAC TCG TTA TGT GGG TGA TGT ACT TTG CAA ATA GTT TCA GAC TTT TCC GAC GTC CTC GAA CTT TTT GGG CAT GGA ATC CTG AGG TTA ATG CAA TCA CTG TCA CAA CCG TGT TGG GAC AGA CAT ACT ATC ACC CCA TTC AAC AAG CTC CAA CAG GCA TTA CTG TGA CCT TGT TGA GCG GCG TGC TTT ACG TTG ACG - 3’

In the standard curve wells, 5ul of dilution and 1 ul water was added per well, replicated 2x vertically. The standard curve for each PMMoV qPCR plate was generated using Quantitative Synthetic Severe Acute Respiratory Syndrome-related Coronavirus RNA obtained from ATCC (Manassas, Virginia). A known quantity of the 229E GBlock was utilized to create the standard curve dilution series, the process of which is described in further detail in the PMMoV section. In the no template control (NTC) wells, 6ul sterile nuclease-free water was added per well, replicated 2x vertically. The plate was then sealed and centrifuged before being put into a Bio-Rad CFX96 Touch Real-Time PCR Detection System thermocycler. The following reverse transcriptase (RT) and qPCR programs were used:

- N1 & N2 Program:
    - RT (Reverse Transcriptase) 1: 10 min at 50C, 1 rep
    - RT 2: 10 min at 95C, 1 rep
    - PCR (Polymerase Chain Reaction) 3: 3 sec at 95C
    - PCR 4: 30 sec at 55C
    - PCR 5: Repeat steps 3 and 4 45x
- 229E Program:
    - RT 1: 10 min at 50C, 1 rep
    - RT 2: 10 min at 95C, 1 rep
    - PCR 3: 15 sec at 95C
    - PCR 4: 1 min at 60C
    - PCR 5: Repeat steps 3 and 4 45x

PMMoV

Each effluent sample was then tested via qPCR a second time for PMMoV, for the purpose of normalization. The following primers and probes were used for this qPCR, obtained from Integrated DNA Technologies (Coralville, IA):

- PMMV-FP1-rev
    - 5’ - GAG TGG TTT GAC CTT AAC GTT TGA - 3’
- PMMV-RP1
    - 5’ - TTG TCG GTT GCA AAT GCA GT - 3’
- PMMV-Probe1
    - 5’ - FAM-CCT ACC GAA GCA AAT G-BHQ1 - 3’

The Bio-Rad CFX96 Touch Real-Time PCR Detection System thermocycler was again utilized.

The standard curve for each PMMoV qPCR plate was generated using the following GBlock sequences (obtained from Integrated DNA Technologies Oligonucleotide Specification Sheets, Coralville, IA)

- Sequence: PMmV-FP1-rev
    - 5’- GAG TGG TTT GAC CTT AAC GTT TGA -3’
- Sequence: PMMV-RP1
    - 5’- TTG TCG GTT GCA ATG CAA GT -3’

The GBlock sequences were first resuspended by adding 125ul of TE buffer. Then, the following calculation for dilution was utilized:

(C1(copies))(V1(ul))=(C2(copies)(V2(ul))

106.58 ul of resuspended gblock and 893.42 ul of TE was aliquoted to obtain the 1.00E+00 dilution. 100 ul of dilution and 900 ul of TE was aliquoted to obtain the 1.00E-01 dilution. This was repeated for the 1.00E-02 and 1.00E-03 dilutions. To perform the standard curve dilution, 100 ul of the 1.00E-04 dilution and 400 ul of TE was aliquoted to obtain the 1.00E-05 dilution. Then, 100 ul of dilution and 900 ul of TE was aliquoted to obtain the 1.00E-06 dilution. This was repeated for the 1.00E-07, 1.00E-08, and 1.00E-09 dilutions.

As in the SARS-CoV-2 qPCR plate, 20ul total volume was aliquoted per well. In each well utilized, 14ul of TaqPath 1-Step RT-qPCR Master Mix, CG (Applied Biosystems, Bedford, MA) was aliquoted. 6ul of sample was added to unknown wells, replicated 3x horizontally. 6ul of positive control solution (PMMoV GBlock) was added to PC wells, replicated 2x vertically. 5ul dilution and 1ul water were added to each standard curve well, replicated 2x vertically. In No Template Control (NTC) wells, 6ul sterile nuclease-free water was added per well, replicated 2x vertically.

The following program was used to run PMMoV RT-qPCR for the effluent samples:

- RT 1: 10 min at 50C, 1 rep
- RT 2: 10 min at 95C, 1 rep
- PCR 3: 5 sec at 95C
- PCR 4: 1 min at 60C
- PCR 5: Repeat steps 3 and 4 45x

Statistical Analysis

Data was log transformed but not normalized. After raw data of log10 viral copies per liter was obtained, several equations were applied to find estimated disease presence. Application of equations to calculate approximate disease prevalence was modeled off of Curtis et al. (2020).

Equation 1 calculates viral load per individual.

$$Load_{indiv} = C_{indiv} \times m$$

$Load_{indiv}$= viral load per individual (copies/day)

C_{indiv}= concentration of virus in fecal matter, i.e. shed rate (copies/g)

m = average mass of feces produced per individual per day (g/day)

Variable m is well established as approximately 128.83 g (Rose et al.). C_{indiv} is approximated as 4170003 copies/g feces ((B. Schmitz, personal communication, Jan. 9, 2022).

Equation 2 calculates total viral load of the wastewater treatment plant.

$$Load_{WWTP} = C_{WWTP} \times Q \times f$$

$Load_{WWTP}$= total viral load to WWTP (copies/day)

C_{WWTP}= concentration of virus in wastewater samples (copies/100mL)

Q= Plant flow (Millions of gallons [MG] per day)

f = conversion factor (100 mL to MG)

Equation 3 utilizes the products of Equations 1 and 2 to estimate disease prevalence.

$$I = \frac{Load_{WWTP}}{Load_{indiv}}$$

I = approximate number of people in WWTP service area infected and at peak shedding rate

In addition to these equations, estimated infection per 10,000 people was calculated.

I / Population of WWTP service area * 10,000

Normalization

Results were normalized to sample population via PMMoV quantification using the method modeled in Fuqing et al. One sample from each month, excluding February due to data limitations, was normalized to achieve a general depiction of the normalized data. A deviation factor determining the deviation of PMMoV copies from the median of PMMoV copies in all samples was calculated for each sample using the following equation:

Deviation factor = 10 ^ [k x (sample CT - median CT)

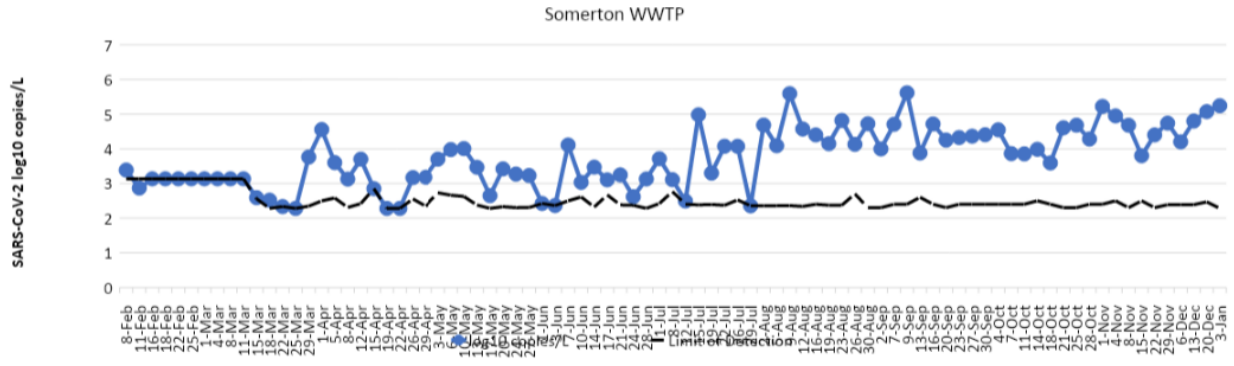
where k is the slope of the standard curve (-3.058) and CT is the cycle threshold (also seen as cycle quantity, represented by CQ). After deviation factor was calculated, the mean of SARS-CoV-2 copies for the month represented by the sample deviation factor was divided by that deviation factor. The equation is as shown:

Normalized viral copies = Non-normalized monthly mean copies / deviation factor

Results

Somerton Wastewater-Based Epidemiology Raw Data

Figure 1



SARS-CoV-2 log10 viral copies/L for all dates sampled between 2/8/21 and 1/3/22. | Viral copies/L is essentially viral load, or the most basic quantification of how much SARS-CoV-2 is in each sample. | Figure created by Yuma Center for Excellence in Desert Agriculture

Out of 61 data points over approximately one year (February 4, 2021 to January 3, 2022), SARS-CoV-2 log10 copies/L was recorded for the Somerton municipality every three to five days. Within that range, there is a clear upward trend over time.

From February to early March, when the limit of detection (LOD) was higher due to lack of certain materials, viral copies remained fairly constant and near the LOD. When the LOD dropped, viral copies at first dropped with it, and then spiked significantly in mid to late March. Another major spike in viral copies peaked in early May. From then on until approximately early August, viral copies continued to fluctuate heavily, increasing and then dropping back near the

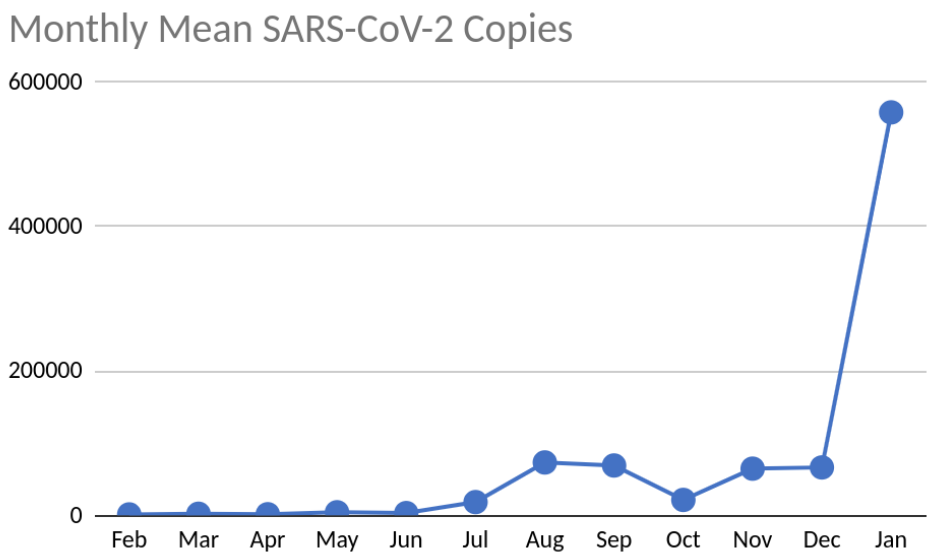
LOD in cycles. Peaks did not seem to be as significant or consistent as the late March/early April and early May spikes.

Beginning in early August, the trend of viral copies no longer continued to drop back near the LOD. Instead, viral copies tended to exhibit a steady rise moreso than drastic fluctuations.

Before mid-July, level of concern primarily stayed at Level 1. However, since that time, it has remained at Level 2 fairly consistently. Level 1 is interpreted as necessitating enhanced awareness for wearing masks and social distancing. Level 2 is interpreted as necessitating coordination with Public Health Services if response action is deemed essential.

Monthly Mean SARS-CoV-2 Copies

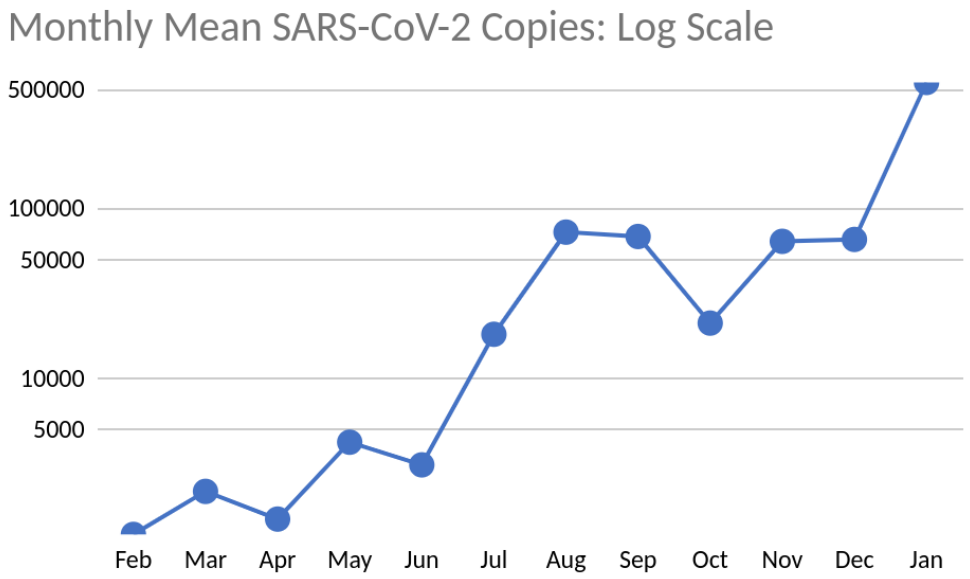
Figure 2



Average viral load, i.e. viral copies/L, per month. | Utilized to decrease noise and increase readability of raw data. Removes outliers and presents a general trendline of viral load progression, and thus, to an extent, pandemic progression.

For a more concise and smooth picture of months' worth of data points, the viral copies/L for each month was averaged as depicted in Figure 2. This allowed for more clear comparison to clinical data given that noise was eliminated. The monthly mean copy values were also log-scaled as shown in Figure 3 to allow for greater detail in visual analysis:

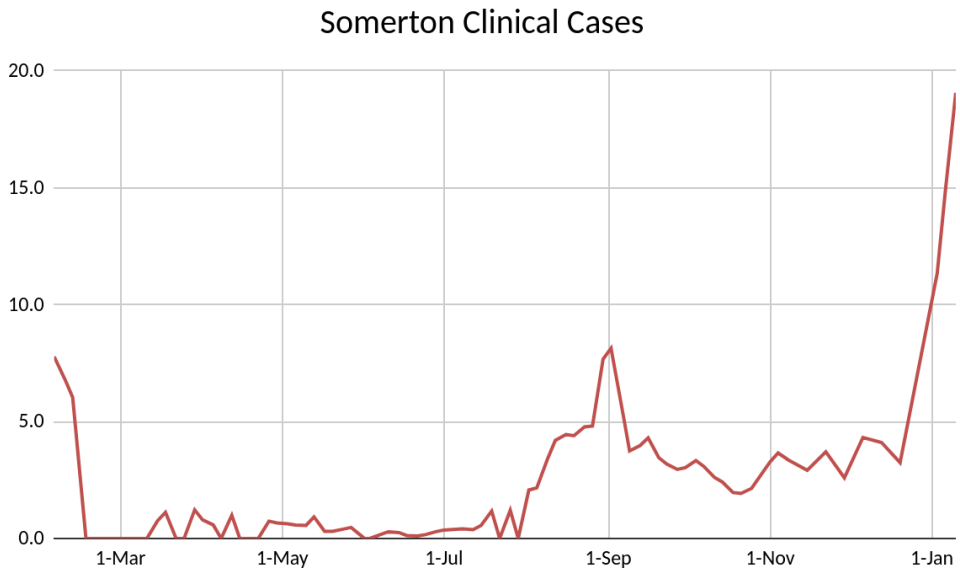
Figure 3



Log-scaled version of Figure 2. | Depicts finer details for ease of visual analysis and comparison

Due to inconsistency in unit of measurement, monthly mean copies and clinical data may not be combined in a single graph. However, the contour of the data may be compared for a general idea of how the two sets differentiate.

Figure 4



Somerton clinical case numbers approximated based off of Yuma clinical case numbers. | Calculated proportional to the population of Somerton, given the similar behavior of the Yuma municipalities.

Figure 4 depicts approximated case numbers for the Somerton municipality of Yuma. Clinical data from the entirety of Yuma had to be adjusted to estimate case numbers for Somerton alone. The following calculation was utilized:

([Cases]*2.13787)*0.0842

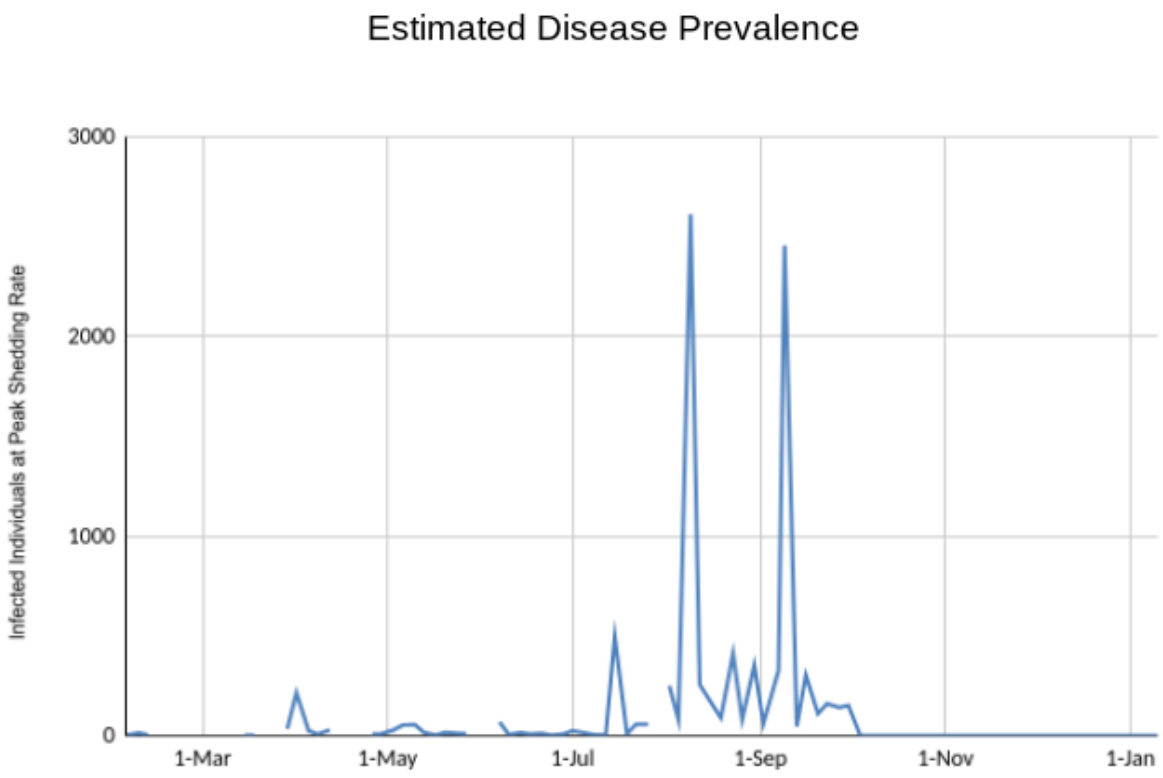
The first constant, 2.13787, was obtained by dividing the current population of Yuma county by 100k, so that cases could first be adjusted for all of Yuma county. The second constant, 0.0842, was obtained by approximating the percentage of Yuma county citizens who are part of Somerton via dividing Somerton's population by Yuma county's population.

Both Figure 3, representing quantity of viral RNA, and Figure 4, representing number of people at peak shedding rate, display a similar general shape. Each data set rises to a first spike, drops down some-- but not as low as prior-- and then soon begins to rapidly rise.

However, while each graph shows a similar progression in the spikes and dips of the virus, the timing differs. In Figure 3, one can observe a minor spike in May preceding the more major spike in August. The same minor, preceding spike appears in Figure 4-- however, it is not recorded until August. The more significant spike that these small peaks lead up to occurs in August in Figure 3, but in September in Figure 4. Finally, the rapid rise that appears to continue at present and reaches each graph's highest point begins in early December in Figure 3, yet not until late December in Figure 4.

Estimated Disease Prevalence: Infected Individuals at Peak Shedding Rate

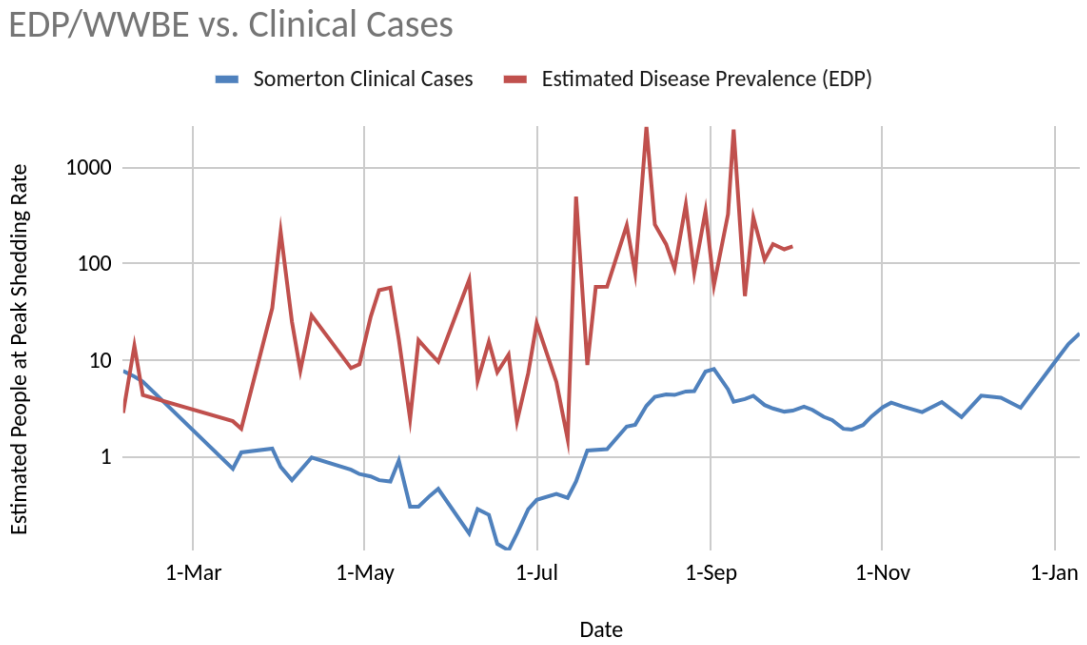
Figure 5



Estimated disease prevalence represents the number of new cases on a given day based on viral RNA quantification in wastewater. | It measures the number of people at peak shedding rate, which occurs in the first 8 days of infection. It can be easily compared to clinical data.

Moving on from raw data based on viral copies, statistical analysis as seen in the Methods section allowed for calculation of estimated disease prevalence (EDP) -- i.e., the approximate amount of individuals at peak shedding rate, which occurs approximately within the first 8 days of infection. As shown in Figure 5, approximated case numbers had a high amount of variation. As a result, EDP was log-scaled. EDP and clinical case numbers were also combined into one graph for ease of comparison.

Figure 6



Log-scaled comparison of Somerton clinical cases and estimated disease prevalence (EDP). | Log-scaled due to dramatic difference in values and EDP's large peaks. EDP data cuts off after November due to lack of flow rate data. Null values are plotted for ease of analysis.

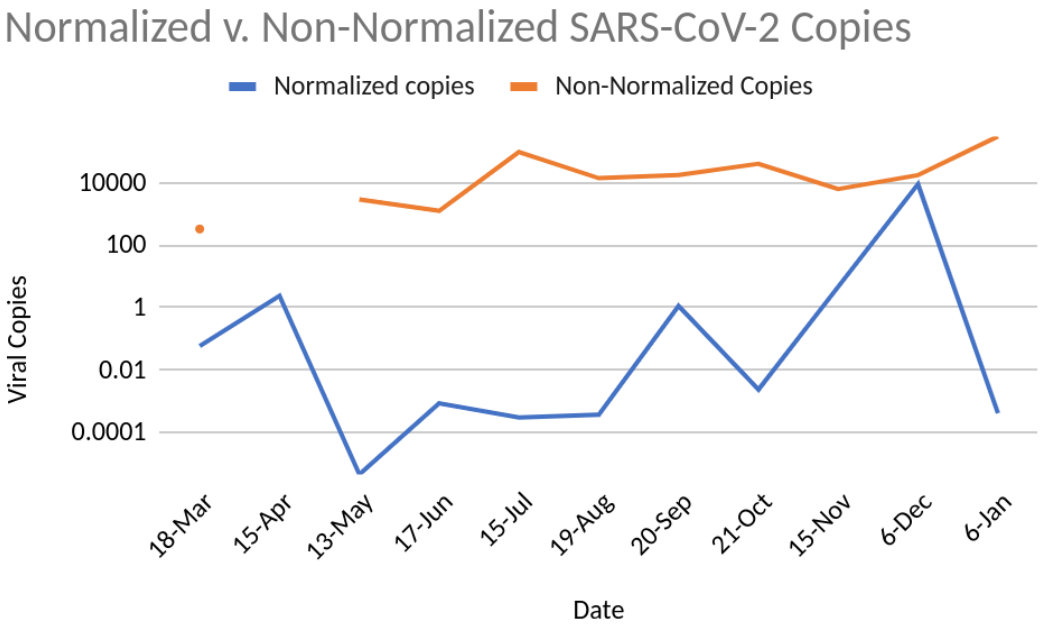
It is immediately apparent that EDP, or cases estimated by WWBE, are significantly higher than clinical numbers. Where wastewater-based cases pass 2500, clinical cases don't even hit 20. However, the general contour of the data, just as in the comparison of clinical cases vs. viral copies, is similar.

To more exactly evaluate the difference between EDP and clinical cases, differences in case numbers fall within a range of -5, one of the two outliers in which clinical data had larger case numbers than WWBE, and 2610. The average difference was that EDP case numbers were 178 greater than clinical case numbers.

PMMoV Normalized Viral Copies

In an attempt to reduce noise and allow for more accurate comparisons between arrays, SARS-CoV-2 viral load data was normalized based on quantification of PMMoV in samples. One sample for each month was normalized to achieve a general picture.

Figure 7



PMMoV-normalized SARS-CoV-2 viral copies vs. raw data. | Gaps in non-normalized data are due to non-detect or lack of flow rate data. Log-scaled for ease of comparison. Only the data points normalized were plotted.

As is clear in Figure 7, normalized values were significantly lower than non-normalized viral copies. Normalized values also appeared to much more variation. Recall Figure 3. The normalized data shows much more dramatic peaks and dips; however, this could be a result of

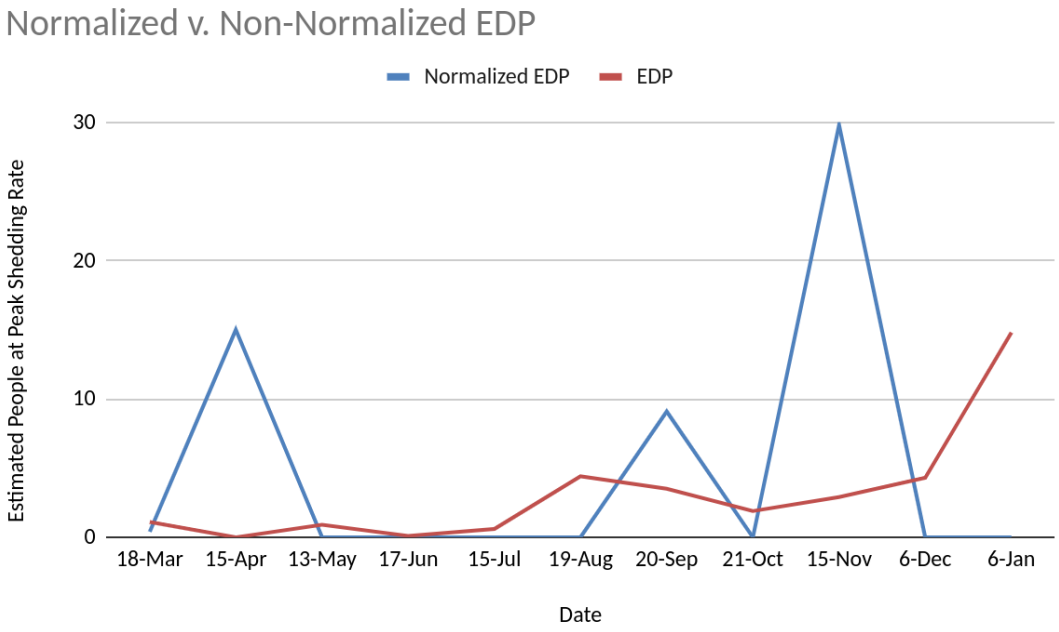
the lesser number of data points. Interestingly, normalized data depicts a spike in April level with one in September. This conflicts with the clearly separate contour of the monthly mean graph. The normalized data also sharply drops down in January, unlike monthly mean or clinical data.

Normalized viral copies may be marginally more aligned with clinical cases. Recall Figure 4. Both graphs show a spike somewhere between late February and April, a spike of similar size in September, and then a higher peak in January. However, there is no sharp drop in clinical cases during January.

Normalized viral copies may even lag behind clinical data-- the September peak occurs early in the month in clinical case numbers, yet late in the month in viral load. In addition, the first spike occurs in late February in clinical cases, but in mid-April for normalized viral copies.

Normalized vs. Non-Normalized EDP

Figure 8



Calculated estimated disease prevalence using normalized SARS-CoV-2 viral copies vs. non-normalized SARS-CoV-2 viral copies.

EDP for normalized viral load was calculated and compared to non-normalized EDP for the same data-points in Figure 8. It is immediately clear that the normalized data shows much more dramatic deviation while the non-normalized data shows a steady rise. Within its 3 peaks, which were its only non-zero points, the normalized EDP approximated a total of 54.3 cases. The non-normalized EDP estimated a total of 34.5 cases-- approximately 20 cases less.

Discussion

Wastewater-based epidemiology surprised researchers with its accuracy in estimating disease prevalence(B. Schmitz & S. Slinski, personal communication, Jan. 12, 2022). In the field of epidemiology, a host of variables can muddle results, to the extent that 25% variability is typically considered very accurate. WWBE has high accuracy given its many unknowns– individual shedding rate, shedding rates of different demographics, exact fecal production per individual, effect of sampling time and date, effect of visitors to the sample area, the many opportunities for contamination in the field. The potential of WWBE for tracking infectious disease cannot be understated. Already, the US government has invested millions of dollars into creating permanent research infrastructure for this subfield.

In this research, two primary questions were put forth– what is the approximate number of unreported cases in the Somerton municipality, and what is the approximate lag time between wastewater and clinical data? These inquiries were tailored to broadly evaluate how WWBE may be used to track SARS-CoV-2.

The CDC estimates 1 in 4 SARS-CoV-2 cases are reported, but claims that this estimate may be low. Based on this data, unreported cases in the Somerton area were hypothesized to total some quantity greater than four times clinical case numbers at any given time. Addressing estimation of lag time was more difficult than looking at CDC statistics– no standard lag time range has yet been established, hence its prioritization in this research. Compiling the ranges estimated in

Peccia et al., Chavarria et al., and Ahmed et al., lag time was hypothesized as somewhere within the range of 6-14 days.

Unreported Cases Hypothesis

As expected, estimated disease prevalence based on WWBE data was much higher than clinical case numbers. All but 2 outliers supported this expectation. The average difference between clinical and WWBE case numbers also helps make clear how well this statement was supported--EDP predicted a mean of 178 cases greater than clinical data. To more specifically reference the numerical hypothesis, each clinical case value was multiplied by four and compared to EDP. Out of 51 data points, only 6 EDP values were less than four times clinical case numbers. I.e., 88% of data gathered supports the hypothesis clinical case numbers are less than a quarter of actual disease presence.

Lag Time Hypothesis

That WWBE detects COVID-19 cases before clinical data is strongly supported by the comparison of monthly mean viral copies and clinical cases. Based on this comparison, there is a clear lag in clinical data's recording of spikes in viral RNA presence, and thus its recording of cases. Three peaks were defined in which lag between data sets was clear. This "lag time" can be

approximated as a range between one and three months. Based on that conclusion, the lag time hypothesis was an underestimation.

However, it must be noted that normalized viral copies conflicted with the lag time hypothesis. Comparison of normalized viral load and clinical data showed an unexpected lag in normalized WWBE data, rather than clinical cases. Reasons for this result are unclear. It is possible that there was some error in the normalization process, or that not enough data points were normalized. Given both these concerns and the inconsistency of normalized viral copies with monthly mean viral copies, the clinical data vs. monthly mean was seen as more reliable data. Thus, it can be concluded that the lag time hypothesis of 6-14 days was a significant underestimate.

Limitations & Errors

Field epidemiology is, unfortunately, plagued by many unknowns. In this particular subfield, the most significant is fecal shedding rate of individuals. Determining the exact number of cases based on viral RNA fecal shedding is challenging, because many correlating factors could influence the viral RNA shed per patient. Individuals may shed different quantities of RNA based on case severity, demographic, progression, and other factors. The constant used for this research was obtained from lead researchers of viral fecal shedding rate who believe they have found a reliable value. The study is not yet publicly available, but its data is consistent with years of research.

Another significant limitation of this research is that many outside variables are impossible to control. From the time that wastewater samples are produced, to when they are collected, to when they are tested, there are countless opportunities for results to be altered by uncontrolled factors. For example, variable m, the amount of feces produced per individual per day, is not consistent or able to be controlled; while it is treated as a constant in this research, it is truly a range. However, despite the many uncontrolled outside factors, the reliability and accuracy of WWBE data compared to other field epidemiology subfields is very high.

There are also limitations in the availability of data point collection locations and ensuring that clinical data is available. Samples are available from many areas both within and outside of the Yuma area, on many scales, making the former issue minor. However, the latter has placed greater limitations-- publicly available data is often only available on certain scales. This requires more approximation than preferable.

Future Research

As more data becomes available concerning fecal shedding rate and clinical numbers become more specific and easily available, conducting such research will provide more accurate results and a more precise understanding of the efficiency of WWBE as a SARS-CoV-2 tracking mechanism. There are many questions still left unanswered in this growing field. WWBE also has the potential to track other infectious diseases– much research focuses on SARS-CoV-2 at this time, but WWBE will likely expand in the future to cover a range of infectious disease

tracking. Future research may address fecal shedding differences by demographic, optimal normalization methods, or other questions that arise as this field grows along with the pandemic.

Acknowledgments

Research was conducted with the support of the STAR Labs program. Director of research Margaret Wilch and facilitator Evy Nguyen were among STAR Labs personnel that contributed assistance and feedback. Dr. Christian Roessler, serving as a mentor, provided guidance and resources throughout the duration of the research. At the Yuma Center of Excellence for Desert Agriculture, laboratory work was conducted under the supervision of Dr. Stephanie Slinski. Dr. Bradley Schmitz provided field-specific direction and resources.

Bibliography

Ahmed, Warish, et al. “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, Elsevier, 18 Apr. 2020, www.sciencedirect.com/science/article/pii/S0048969720322816.

Betancourt, et al. “Covid-19 Containment on a College Campus via Wastewater-Based Epidemiology, Targeted Clinical Testing and an Intervention.” PubMed, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/33743467/.

“CDC Covid Data Tracker.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, covid.cdc.gov/covid-data-tracker/#county-view?list_select_state=Arizona&data-type=Risk&list_select_county=4027.

Chavarria-Miró, et al. “Time Evolution of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV-2) in Wastewater during the First Pandemic Wave of Covid-19 in the Metropolitan Area of Barcelona, Spain.” Applied and Environmental Microbiology, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/33483313/.

Curtis, Kyle, et al. “Wastewater SARS-COV-2 Concentration and Loading Variability from Grab and 24-Hour Composite Samples.” MedRxiv, Cold Spring Harbor Laboratory Press, 1 Jan. 2020, www.medrxiv.org/content/10.1101/2020.07.10.20150607v1.

D'Aoust, Patrick M., et al. “Quantitative Analysis of SARS-COV-2 RNA from Wastewater Solids in Communities with Low COVID-19 Incidence and Prevalence.” Water Research, Pergamon, 23 Oct. 2020, www.sciencedirect.com/science/article/pii/S0043135420310952.

“Data Analysis -- Normalization.” Cell Biology Protocols, www.sciencegateway.org/protocols/cellbio/microarray/normal.htm.

Feng, et al. “Evaluation of Sampling, Analysis, and Normalization Methods for SARS-COV-2 Concentrations in Wastewater to Assess COVID-19 Burdens in Wisconsin Communities.” ACS Publications, pubs.acs.org/doi/10.1021/acsestwater.1c00160.

Fuqing, et al. “SARS-CoV-2 Titers in Wastewater Are Higher than Expected from Clinically Confirmed Cases.” ASM Journals, https://journals.asm.org/doi/10.1128/mSystems.00614-20#fig4

Medema, et al. “Presence of SARS-Coronavirus-2 RNA in Sewage and Correlation with Reported COVID-19 Prevalence in the Early Stage of the Epidemic in the Netherlands.” ACS Publications, pubs.acs.org/doi/10.1021/acs.estlett.0c00357.

Peccia, Jordan, et al. “Measurement of SARS-COV-2 RNA in Wastewater Tracks Community Infection Dynamics.” Nature News, Nature Publishing Group, 18 Sept. 2020, www.nature.com/articles/s41587-020-0684-z.

Rose, et al. “The Characterization of Feces and Urine: A Review of the Literature to Inform Advanced Treatment Technology.” Critical Reviews in Environmental Science and Technology, U.S. National Library of Medicine, pubmed.ncbi.nlm.nih.gov/26246784/.

"U.S. Covid Risk & Vaccine Tracker." Covid Act Now, covidactnow.org/us/arizona-az/county/yuma_county/?s=28101696.

"Yuma County 2022 COVID-19 Daily Updates." Yuma County AZ, www.yumacountyaz.gov/government/health-district/divisions/emergency-preparedness-program/coronavirus-2019-covid-19-yuma-county-updates/yuma-county-updates.