# Introduction

In December 2019, a virus known as SARS-Cov-2 (COVID-19) was identified in Wuhan, China. This new variant of the SARS coronavirus caused a worldwide pandemic with far-reaching consequences. An important factor contributing to the silent spread of the virus is the fact that 20 to 40% of the patients show no symptoms, (Vallejo et al., 2019).

Although patients do not always show symptoms of the virus, they often do excrete RNA particles of the virus in their faeces as shown by for example Pan et al. (2020). The virus can sustain itself for a long period within the faeces, in some cases even for one or more months after the respective patient has tested negative for COVID-19 in their respiratory samples (Vallejo et al., 2019). Therefore, the amount of RNA particles could be a valuable indicator of the true number of COVID-19 patients within a country or municipality.

The aim of this project is to estimate the true number of infected people in the Netherlands based on amount of RNA particles in sewage water.

**Structure of the data**

We are working with two datasets, both collected by the Dutch National Institute for Public Health and the Environment (RIVM). The first dataset contains the total number of positive tests reported per day per municipality. It also contains information on key characteristics of the municipality, such as population density and which security region it is part of. The second dataset contains data recorded on the level of sewage treatment plants. The key variable here is the average concentration of SARS-CoV-2 RNA measured in the daily amount of sewage water per 100.000 inhabitants. This dataset also contains crucial metadata of the sewage treatment plants, such as in which security region the area of responsibility of this treatment plant falls. The two datasets were matched to each other by the variable in which security region a municipality and a treatment plant’s area fall respectively.

**Challenges in the data**

Our goal is to estimate the true number of COVID-19 patients at any given day, based on the data we have available describing the RNA flow in the sewage water. To do this, we first have to establish if there is a relationship between the RNA flow and the number of positive tests per day. Given the data structure, we are faced with a few challenges before we can take on this question.

First of all, the data on the number of positive tests are recorded on the level of municipalities, whereas the data on RNA flow are recorded on the treatment plant level. In the most straightforward cases, we can aggregate the RNA flow data to municipality level data by virtue that the datasets were already matched by security region code. See SR1 of Figure 1.

However, some sewage treatment plants also treat water from outside their primary security region, creating double entries in the dataset and making simple matching impossible. Luckily, the dataset also provides information on what percentage of the water a treatment plant processes comes from which security region, so we can weight the RNA flow by this variable. See SR2 of Figure 1.

Furthermore, some very large municipalities produce so much sewage water in one day, that multiple treatment plants are required to process it, causing a second kind of double entries. See SR3 of Figure 1. Unfortunately, there are no data available that specify how much of the water from these large municipalities goes to which treatment plant, making it impossible to establish if there is a relationship between the RNA flow and the number of positive cases on the municipality level.

We take the above relationships into account in our analyses. Furthermore, we conclude that we have to aggregate our data to the Security Region level when establishing a relationship between RNA flow and number of positive cases, to get an accurate indicator. Lastly, the data we use in our analyses are a subset of the original dataset, because the dataset only includes all sewage treatment plants from the 7th of September onwards.

### Methods and preliminary results

We work towards our goal through asking and answering several sub-questions. The following section contains the analyses and preliminary results of these sub-questions.

### How does the total number of reported infections in the Netherlands as a whole develop over time?

We wrote a function to aggregate the data by the number of reported infections in a municipality on a given day to the total number of reported infections on that day in the Netherlands as a whole, discarding double entries per municipality. The resulting output is visualized in Figure 2a, where you can see the total number of reported infections we have had over time in the Netherlands. The second wave is also clearly visible, where you see the total number of reported infections increasing more rapidly in October compared to September.

We are also interested in the daily fluctuations of the number of reported infections in the Netherlands. To show this, we altered the function such that it would calculate the difference in total reported cases between each day and the day before. The results are visualized in Figure 2b.

### What does this trend look like displayed per 100,000 inhabitants of the Netherlands?

Next, we visualized the number of reported cases per 100,000 inhabitants of the Netherlands. Again, we decided to do this for both the total number of infections we have had over time, as well as the daily differences. As you can see in Figures 3a and 3b, the shape of the distribution is the same, but the scale on the y-axis is different. This is expected, as we took the functions of the previous step and multiplied the total number of reported infections by the number of inhabitants of the municipality divided by 100,000.

### What is the mean level of RNA particles found in the water per 100,000 inhabitants?

We saw that the number of infections had been increasing since the beginning of September and peaked in October. One would expect the amount of RNA particles in the sewage water to increase with the number of infections. In Figure 4, the mean number of RNA particles per 100,000 inhabitants summed over all sewage water installations for every day is plotted. The shape of the plot loosely follows the same trend as the plot of the reported infections, but its shape is much less smooth. This could be due to the fact that not all treatment plants report the RNA flow every day, and that the number of reporting treatment plants differs per day.

### What is the relationship between RNA particles and the total number of infections on one given moment?

As mentioned in the data description, some installations process water from multiple security regions, rendering the current RNA flow variable unrepresentative. We do have a variable available that contains the proportion of water from the security region processed by that respective installation. We multiplied these two columns to create a representative RNA flow variable.

Now that the RNA flow has been weighted, we want to inspect how the RNA flow is related to the increase in the number of infections on the level of the security regions. To achieve this, we altered the current functions for calculating the RNA flow and increase in number of infections per day so that these values are reported for each security region per day. Subsequently, we compared the RNA flow from one day to the increase in the number of reported infections from seven days later. The reason for this is that there is generally a seven-day time lag between a person contracting COVID – leaving RNA particles in the sewage water – and getting a positive test (Peccia et al., 2020). A scatterplot with the aforementioned data should provide a first impression of how they are related to each other. Since not every security region has data about the RNA flow on any given day and data about the increase in the number of infections seven days later, only security regions that have data for both days are selected. Subsequently, the correlation for those data is calculated.

Figure 5 shows the correlation between the RNA flow on the 8th of September and the increase in the number of infections on the 15th of September. As can been seen from the scatterplot, these data do not seem be strongly related to each other. Additionally, the correlation associated with this scatterplot is 0.027, which indicates that these data almost show no relationship to each other. This strikes us as very surprising, since one would expect a relationship between the RNA flow and the increase in the number of infections. In the next step, we extend the analysis conducted over the period of September until November to see if this trend persists.

### How does this relationship look over time?

We calculated the correlation between the RNA flow on the first day and the increase in infections seven days later for each available combination of days (i.e. available combination of one date and the date seven days later). Finally, we plotted all the correlations over the period in a scatterplot.

Figure 6 shows that the correlations fluctuate over time, which seems very interesting. If the data points would have formed a straight horizontal line, the estimation of the true number of infections in the Netherlands would have been more straightforward, since it would mean that the relationship between the RNA flow and the increase in infections would have been the same for every day. The mean correlation between the datapoints in this scatterplot is 0.33, which is a moderate positive correlation. This means that as the pandemic progresses, the relationship between the RNA flow and the increase in the number of infections becomes stronger. We have two hypotheses as to why this would be the case: it could be a result of increased consistency and precision in the measurements as we get more rigorous in reporting these data, or – as the number of cases has only increased over the period we analysed – it could mean that RNA flow becomes a better predictor of the total number of reported cases, as both

### Preliminary conclusions

As of yet, we are unable to make a reliable estimate of the true number of infected people in the Netherlands. First and foremost, this is due to the unstable nature of the correlation between the RNA flow and the total number of recorded infections. However, the moderate, positive correlation between these two variables indicates that the RNA flow becomes an increasingly better predictor as we get further along in the pandemic. If we do essay to construct a prediction model, we recommend to use only the most recent RNA flow data to build it, as they are likely the most reliable.

Our next line of investigation will be to try to combine the RNA flow as a predictor with other available variables in our dataset, such as population density. We would also like to separate this analysis by security region, to isolate the effect of geographic location on the predictor.

If we had infinite time and resources, we would like to enrich our analysis by incorporating other types of data in our analysis. For example, we could use behavioural data on when people are more or less likely to get tested when displaying symptoms, to improve the estimate of the true number of infected people. Or environmental data on how much rain fell in each region on each day, to correct for the influence this might have on the concentration of RNA particles in the sewage water. For now, we conclude that the RNA flow is an imperfect indicator of the true number of infected people in the Netherlands, but that it could potentially be very valuable in combination with other data sources.

### References

Pan, Y., Zhang, D., Yang, P., Poon, L. L. M. & Wang, Q. Viral load of SARSCoV-2 in clinical samples. The Lancet. Infectious diseases 20, 411-412, doi:10.1016/s1473-3099(20)30113-4 (2020).

Peccia, J., Zulli, A., Brackney, D.E. et al. Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. Nat Biotechnol 38, 1164–1167 (2020). <https://doi.org/10.1038/s41587-020-0684-z>

Vallejo, J. A., Rumbo-Feal, S., Conde-Pérez, K., López-Oriona, Á., Tarrío, J., Reif, R., ... & Veiga, M. C. (2020). Highly predictive regression model of active cases of COVID-19 in a population by screening wastewater viral load. medRxiv.