scientific reports

Check for updates

OPEN

Statistical analysis of three data sources for Covid-19 monitoring in Rhineland-Palatinate, Germany

Maximilian Pilz✉, Karl-Heinz Küfer, Jan Mohring, Johanna Münch, Jarosław Wlazło & Neele Leithäuser

In Rhineland-Palatinate, Germany, a system of three data sources has been established to track the Covid-19 pandemic. These sources are the number of Covid-19-related hospitalizations, the Covid-19 genecopies in wastewater, and the prevalence derived from a cohort study. This paper presents an extensive comparison of these parameters. It is investigated whether wastewater data and a cohort study can be valid surrogate parameters for the number of hospitalizations and thus serve as predictors for coming Covid-19 waves. We observe that this is possible in general for the cohort study prevalence, while the wastewater data suffer from a too large variability to make quantitative predictions by a purely data-driven approach. However, the wastewater data and the cohort study prevalence are able to detect hospitalizations waves in a qualitative manner. Furthermore, a detailed comparison of different normalization techniques of wastewater data is provided.

Wastewater surveillance is an important tool to investigate the current status of the Covid-19 pandemic[1] and has the potential to be used for further viruses[2]. The course of the Covid-19 pandemic is investigated in many countries by wastewater surveillance. Among others, USA[3], New Zealand[4], the Netherlands[5] and Switzerland[6] used this tool as a means of monitoring the detected genecopies in their sewage systems. Among the intended advantages of wastewater monitoring are its independence of individual testing which varies with political guidelines, its comparably low costs, its anonymous character, and that there is no additional effort for the population. However, there are multiple sources of uncertainties that impact the amount of measured genecopies[7].

A high viral load concentration in the communal wastwater does not necessarily mean a major public health burden. The actual relevant quantity is the number of persons who have to go to hospital because of the virus. However, this quantity is no longer recorded systematically in Rhineland-Palatinate as Covid-19 tests in hospitals have mainly stopped[8] and the pandemic will be more and more monitored via wastewater[9]. Before, the rate of infected, hospitalized patients was considered a very important measure of the burden of disease in the population[10,11]. Furthermore, hospitals were regarded as a source of infection and regular testing of all employees and patients was considered as a means of containment[12]. Given the unavailability of this information in the future, the primary objective of this publication is to examine the potential association between the viral load of SARS-CoV-2 in municipal wastewater of Rhineland-Palatinate and the hospitalization count of infected individuals during the period when both measurements were accessible.

In addition to wastewater surveillance, the German federal state Rhineland-Palatinate established a voluntary cohort study in which citizens regularly tested themselves for Covid-19 and reported the results to the clinical trial organization team. Contrary to the officially reported case data, which is biased by testing resources and motivation[13,14], this study data is much more representative and bypasses the problem of a large number of undetected cases.

In this paper, we compare the results of the data sources wastewater surveillance, cohort study, and the reported number of Covid-19-related hospitalizations during their pairwise availability. The wastewater surveillance data are available from December 2022 to May 2023, the cohort study results from January 2023 to May 2023, and the hospitalization numbers from December 2022 to April 2023. Wastewater data is commonly normalized to enhance comparability across different time periods and treatment plants. However, a universally accepted technique for normalization has not yet been established[1,5,15–17]. We illustrate the effect of three common normalization techniques (copy rate, concentration, relative to the reference virus PMMoV) in wastewater and investigate whether a data-based prediction from wastewater surveillance could be an early-warning system for the detection of pandemic waves. While there are many publications that compare wastewater data to officially

Fraunhofer Institute for Industrial Mathematics, Kaiserslautern, Germany. ✉email: maximilian.pilz@itwm.fraunhofer.de

reported test data[18], to the best of our knowledge, to date, there are no comparative results from wastewater and large population cohorts.

In "Methods", we present the three data sources in more detail and describe the applied statistical methodology. In "Results", the results are presented. We conclude with a discussion in "Discussion".

Methods

Data sources

Wastewater

Wastewater was taken twice per week in 15 sewage plants in Rhineland-Palatinate by taking 24-h composite samples, following the same guidelines as were developed for the EU project ESI-CorA[19]. Wastewater was collected time-, volume-, or flow-proportionate and prior to or after the sand filter, depending on the sewage plant. The samples were stored at +4°C to +10°C and processed within 48 h after collection. The laboratory followed the manufacturer's standard protocol for the Promega Maxwell (R) RSC Enviro Total Nucleic Acid Kit for extracting the viral information.

For each measurement, the number of N1 and N2 genecopies, respectively, was determined, which are the two gene targets of the SARS-CoV-2 virus that are commonly used in laboratory testing for Covid-19[1]. In addition, a reference virus, the Pepper mild mottle virus (PMMoV), was measured. This is a plant RNA virus that is known to be a good reference in wastewater to be compared with other viruses[20].

Since wastewater occurrence of viruses may vary with the rainfall, the water volume of each measurement day at each sewage plant was measured. In addition, some wastewater related parameters were collected to detect anomalies, concretely the water temperature, the pH value in water, the chemical oxygen demand, the water conductivity, and total organic carbon. The respective sewage plants together with the number of connected inhabitants and the plants' dry-weather flows are also available. To investigate possible weather-related influences, we added information on the air temperature and the precipitation that is available online by the German Meteorological Service[21]. Since these parameters can not be aggregated on federal state level, they are only used in "Finding 4: Wastewater data alone does not allow quantitative prediction of the cohort study prevalence".

While the wastewater surveillance in Rhineland-Palatinate is still ongoing, we use its measurement values between December 2022 and April 2023, thus for 5 months, corresponding to the availability of the other data sources. We have this data for each treatment facility individually. A publicly available extract of this data is available online[22]. The participating sites include the five cities that were selected for the cohort study.

Cohort study

In 2022 and 2023, the University Medical Center of the Johannes Gutenberg-University Mainz conducted a cohort study ("SentiSurv") on behalf of the Ministry of Science and Health of Rhineland-Palatinate[23]. A cohort was drawn randomly from the five largest cities in Rhineland-Palatinate, namely Mainz, Ludwigshafen, Koblenz, Trier, and Kaiserslautern. These cities are uniformly distributed across the federal state. The participants are almost representative in terms of gender and age except for children who have been excluded for legal reasons. During the course of the first months, all five cohorts reached their targeted size of 2800 volunteers each. In addition to the regular completion of questionnaires, participants performed a rapid test on two fixed days each week and reported the outcome in a mobile app. Starting from January 2023, the number of participants included in the study was large enough to allow a statistically sound interpretation.

The results of the cohort study are available between January 2023 and April 2023, thus for 4 months. We have this data for each of the tested cities individually. The data has been displayed online during the course of the study[24].

Hospitalizations

As an official reference value, the number of Covid-19 related hospitalizations are used. Those are the number of hospitalized persons that were tested positive on Covid-19 for a given day. Note that this differs from the commonly reported number of newly admissioned patients with a positive test. A Covid-19 test was mandatory in hospitals in Rhineland-Palatinate until March 31, 2023, and therefore, the hospitalizations are available between December 2022 and March 2023, thus for 4 months. The data was collected on a weekdaily basis from the Ministerium für Wissenschaft und Gesundheit (Ministry of Science and Health) in Rhineland-Palatinate by calling the individual hospitals and ask for their positively tested patients. We only have these values for the whole of Rhineland-Palatinate, not broken down to cities. To the best of our knowledge, the numbers are not officially reported.

Rhineland-Palatinate

Rhineland-Palatinate is a federal state in the southwest of Germany. It has about 4 million inhabitants and five cities (Mainz, Ludwigshafen, Koblenz, Trier, Kaiserslautern) with more than 100,000 inhabitants. Those are the five cities in which the cohort study was conducted. The largest city is the state capital Mainz with around 220,000 inhabitants. In Rhineland-Palatinate, there are about 100 hospitals that contributed to the number of Covid-19 related hospitalizations described in "Hospitalizations".

Figure 1 shows a map of Rhineland-Palatinate and its location in Germany, which illustrates the distribution of the 15 sewage plants and the five cohort study cities. Furthermore, the map is shaded relative to the population density of the municipalities, ranging from 36 to 2226 people per square kilometer.

Figure 1. Location of Rhineland-Palatinate in Germany and distribution of sewage plants (blue dots), cohort study cities (orange stars), and population density (grey shading). We created the maps ourselves with the python package geopandas[25]. The maps are both based upon open data sources, which we retrieved from http://opendatalab.de/projects/geojson-utilities/ which itself uses geodata from the German Federal Agency for Cartography and Geodesy https://gdz.bkg.bund.de/ and population data from the Federal Statistical Office of Germany https://www.destatis.de/DE/Themen/Laender-Regionen/Regionales/Gemeindeverzeichnis/_inhalt.html. The coordinates from the referenced cities were manually collected from Google Maps.

Data analysis

Parameter computation

From the collected parameters, some additional values could be derived. To appropriately depict the variation in N1 and N2 genecopies simultaneously, one may report their mean. Furthermore, for all three genecopies (N1, N2, and their mean), one can report their absolute value, their value per ml (i.e. normalized by the flow volume adjusted by the sewage plant's dry-weather flow), and their value in relation to the reference virus PMMoV. The latter value is computed as genecopies / PMMoV · 100,000. Three genecopies and three normalization techniques result in a number of nine parameters that can be computed for the wastewater.

The nine numbers defined above (N1, N2, and mean, reported as absolute value, per ml, and per PMMoV, respectively) have to be normalized with the number of inhabitants in order to combine the values per sewage plant to summary values for the whole of Rhineland-Palatinate. For the three genecopies (N1, N2, and mean), this normalization happens in the same manner. Therefore, in the following formulas, the normalization is presented for the absolute value of genecopies, for the genecopies per ml, and the genecopies per PMMoV, where genecopies stands for N1, N2, or their mean. Since the total number of genecopies is an absolute value, one can compute

$$\text{Number of genecopies} = \frac{\sum_{i:\,\text{sewage plant}} \text{genecopies at } i}{\sum_{i:\,\text{sewage plant}} \text{inhabitants connected to } i}.$$

For the relative parameters, the numerator in the above formula would not necessarily increase with increasing number of inhabitants. This implies that sewage plants with small population may be over-represented. Therefore, for these parameters, the following formulas hold:

$$\text{Genecopies/ml} = \frac{\sum_{i:\text{ sewage plant}} (\text{genecopies/ml at } i) \cdot (\text{inhabitants connected to } i)}{\sum_{i:\text{ sewage plant}} \text{inhabitants connected to } i},$$

$$\text{Genecopies/PMMoV} = \frac{\sum_{i:\text{ sewage plant}} (\text{genecopies/PMMoV at } i) \cdot (\text{inhabitants connected to } i)}{\sum_{i:\text{ sewage plant}} \text{inhabitants connected to } i}.$$

The resulting values are multiplied by 100,000 in order to report the values per 100,000 inhabitants.

For the cohort study, the daily prevalence is reported. This means that the number of positively tested participants is divided by the total number of participants. The resulting value is multiplied by 100,000 to obtain the prevalence per 100,000 inhabitants. The cohort study Covid tests were done on Sunday and Wednesday. We denote the Sunday measurement as first measurement and the Wednesday measurement as second measurement of a week.

In order to match the hospitalization data to the two measurements per week, the mean of the number of hospitalizations of Monday, and Tuesday for week measurement one and of Wednesday, Thursday, and Friday for week measurement two are computed, respectively.

Statistics

The resulting parameters can be depicted as point plots. To measure the uncertainty in the data, confidence intervals for all values are needed. The prevalence can be interpreted as a rate and, therefore, Wilson confidence intervals for rates[26] can be computed and multiplied by 100,000. For continuous parameters, we compute the confidence interval for the mean by assuming a normal distribution. To this end, the standard deviation of the measurements has to be estimated. We do this applying a time-shifted concept by computing for each measurement the standard deviation of the respective measurement, and the two measurements before and after this measurement. For the hospitalizations, multiple values are summarized to compute one value per measurement (cf. "Parameter computation"). These values are used to compute the standard deviation for each measurement. Since the reported values cannot be negative, the lower bounds of the confidence intervals are cut at zero.

To illustrate time trends, a smoothing method is needed. We applied locally estimated scatterplot smoothing (LOESS) regression. This method fits for each data point a linear regression in a pre-defined neighborhood, i.e. a certain proportion of the entire dataset around the data point, of this point and predicts the point by this linear regression. By this neighborhood approach, a new linear model is fitted for each data point and thus a local smoothing is performed. The proportion of data points from the entire dataset that is used for the local linear regression is called span. In this paper, we used a span of 50% for the wastewater data and 60% for the comparison of all data due to the lower number of datapoints for the cohort study and hospitalization.

In order to investigate if there is a time shift between different parameters, time lag correlations are computed between the raw values of these parameters. This means that one of the two parameters to be compared is shifted by a time step of l and the correlation between the two resulting time series is computed. The result is the correlation with time lag l. This approach will be used to compare if there is a time shift between prevalence, hospitalizations, and wastewater data.

We tried to use the available data to build prediction models in a data-driven manner, i.e. without the knowledge of any biological background. First, we investigated whether one can predict the number of hospitalizations by wastewater data and the cohort study prevalence. Second, we analyzed the predictive capability of the wastewater data to predict the cohort study prevalence. To this end, regression models were applied. We fitted random forests[27] to predict the outcome of interest. The number of trees in each forest was set to 5000 and the candidates at each split were set to 5. In order to avoid overfitting, leave-one-out cross-validation was applied. This means that for each measurement, a random forest was fitted on all other data points, but not this measurement, and the measurement was then predicted by the resulting random forest.

To make the scenario realistic, past values of the variables were included as covariates into the regression models. Those past values are denoted with lag i when talking about the timepoint i measurements before the current observation. Feature importance was computed by a random forest fit on the full datasets to investigate which covariates influence the prediction remarkably. This feature importance was calculated as the factor by which the random forests prediction error increases when the respective feature is shuffled within the underlying dataset. As error measure, the robust mean squared error (RMSE) was used.

Data was analyzed and visualized using the statistical software R[28], version 4.3.0, and the `tidyverse` packages[29]. LOESS regression was done using R's standard function 'loess'. Random forests were fit with the R package `randomForest`[30] and the R package `iml`[31] was applied for the calculation of feature importance.

Results

We present our results as four main findings that are given as the respective subchapter headings.

Finding 1: Trends in wastewater are present irrespective of genecopy type and normalization technique

Figure 2 shows the three genecopies (N1, N2, and their mean) together with three normalization techniques (absolute value, per ml, and per PMMoV).

Regarding trends, all nine panels show the same pattern with a small wave in late 2022 and a second, larger, wave starting in February 2023. The N2 values are larger than the N1 values but show the same pattern. Normalizing the data by the flow or the reference virus compresses the data without changing the depicted trends. Therefore, normalizing the data seems to act as regularization. When normalizing the data by the flow or the

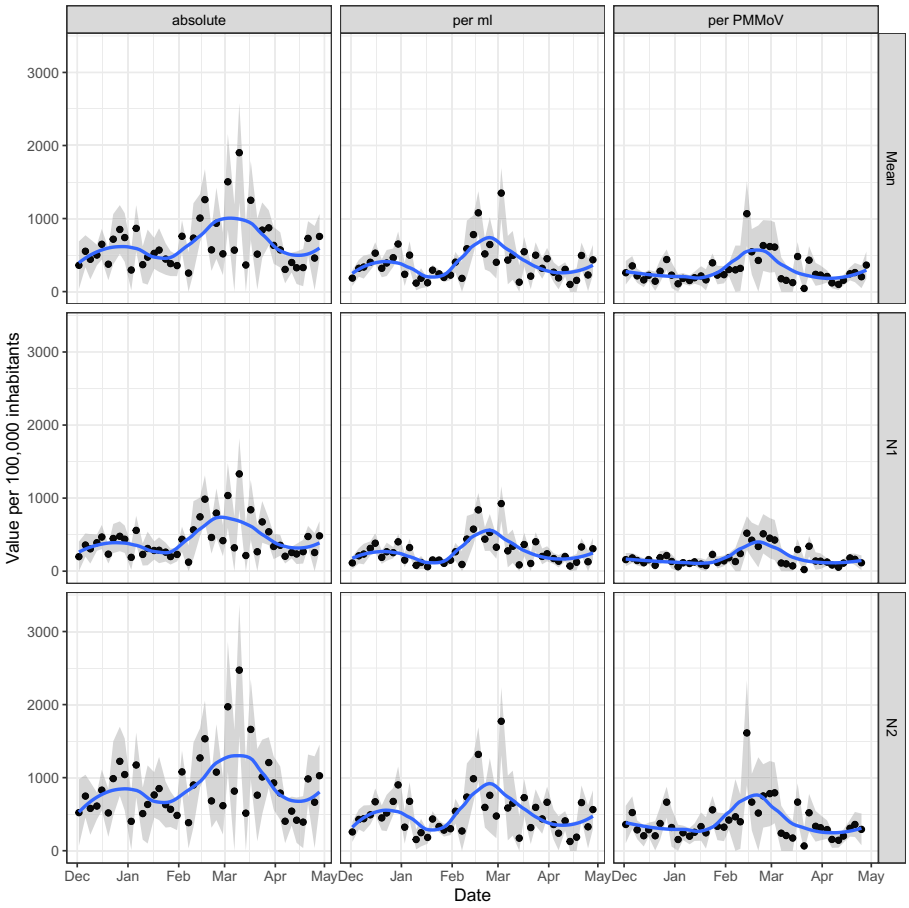

Figure 2. Wastewater values (mean, N1, and N2) in dependence of normalization technique. Main trends are detectable in all types of genecopies and all normalization techniques. Normalizing by flow or PMMoV regularizes the curve.

PMMoV, the second wave reaches its peak before March 2023 while the absolute virus value continues increasing until mid March 2023. This behavior is observed for N1, N2, and their mean analogously. After the peak, all curves decrease and show a moderate increase in the last measurements.

Finding 2: Qualitative trends are observed in all three data sources

Figure 3 compares the three data sources genecopies (normalized per ml), prevalence from the cohort study, and number of hospitalizations according to their availability described in "Data sources".

There are two hospitalization waves. The first, slightly higher wave, has its peak around New Year. After a low point in late January, the second wave has its peak in the first half of March. Due to missing data from 2022, the cohort study prevalence only captures the second wave. The peak of this wave is reached about 7 days earlier than in the hospitalizations. This indicates that the cohort study may be a good tool to detect hospitalization waves with some time lead. The wastewater data capture both waves but in different intensity. The second peak is here distinctly higher than the first one. Of note, the wastewater values show a much larger variance than the other two data sources. This complicates the interpretation of single values and the early detection of potential virus waves.

Finding 3: Cohort study prevalence and wastewater data allow prediction of hospitalizations

Figure 3 indicates that hospitalization waves may be detectable earlier in the wastewater and the cohort study prevalence, respectively. Table 1 shows the time lag correlation between the raw values of the three data sources (genecopies/ml in wastewater, cohort study prevalence, and number of hospitalizations). In each column, two of the sources are compared pairwise by computing their Pearson correlation under the given time lag. The time lag is defined such that the first mentioned value is shifted to the left by the respective time lag.

The genecopies per ml show a slight correlation with the number of hospitalizations 14 to 3 days (4 to 1 measurements) later. The highest correlation values are observed between the cohort study prevalence and the hospitalizations within a time difference of 0 to 7 days (0 to 1 measurements). Genecopies per ml and the prevalence are also positively correlated, the prevalence seems to fit better to the hospitalizations than the genecopies per ml.

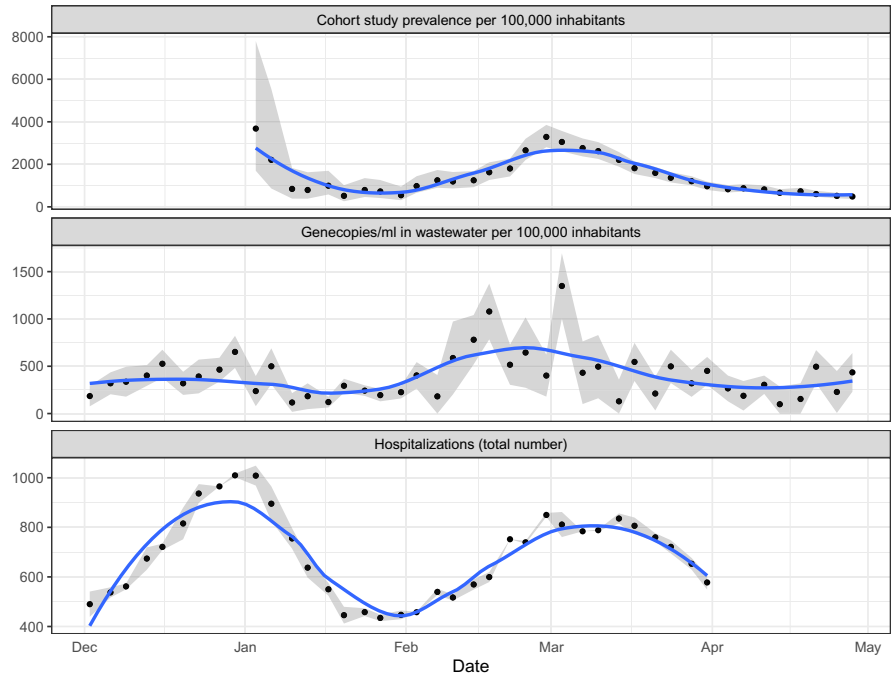

Figure 3. Comparison of prevalence, genecopies, and hospitalizations. The main wave is observable in all three parameters.

Time lag	Gencopies/ml vs. Hospitalizations	Prevalence vs. Hospitalizations	Gencopies/ml vs. Prevalence
− 4	0.382	0.292	0.468
− 3	0.481	0.441	0.591
− 2	0.412	0.572	0.613
− 1	0.370	0.704	0.581
0	0.219	0.765	0.445
1	0.140	0.629	0.419
2	− 0.079	0.415	0.177
3	− 0.192	0.198	− 0.005
4	− 0.397	−0.046	− 0.125

Table 1. Pairwise time lag correlation between the data sources. Each column shows a pairwise comparison of two of the three sources. In the rows, the correlations for the respective time lag are given. Time lag is given in measurements.

Additionally, we investigated whether the number of hospitalizations can be predicted by the cohort study prevalence, raw values of the genecopies, and the genecopies per ml including the past three values of these parameters. As described in "Statistics", random forests with leave-one-out cross-validation were fitted.

Figure 4 shows the results of the hospitalization prediction. In the left panel, the true numbers of hospitalizations are plotted together with the random-forest-predicted hospitalizations. The prediction is able to detect the wave in the first weeks of March, including the increase before and the decrease after the peak. The predicted values are slightly compressed compared with the true values. On the right panel, the feature importances are shown. The three most important parameters are the cohort study prevalences zero to two measurements (thus 0 to 10 days) before the current measurement. This corresponds with the findings of Fig. 3 and Table 1 that there is a certain time shift between the prevalence and the number of hospitalizations. The genecopies in wastewater do not influence the outcome considerably.

Finding 4: Wastewater data alone does not allow quantitative prediction of the cohort study prevalence

As observed in "Finding 3: Cohort study prevalence and wastewater data allow prediction of hospitalizations", the cohort study prevalence is a good predictor of the number of hospitalizations and, therefore, it seems to be a good indicator to investigate the development of the pandemic. Since it is less effort to test the wastewater than

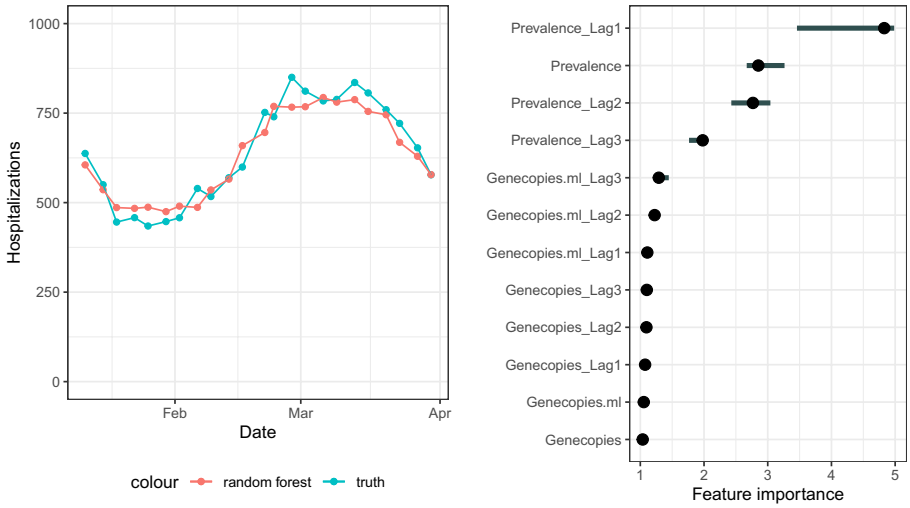

Figure 4. Hospitalization prediction and corresponding feature importance derived by a random forest. The number of hospitalizations can be predicted well. The most important predictors are the cohort study prevalences. The term 'lag' defines the measurement 'lag' before the current measurement.

conducting an ongoing cohort study, it is of interest whether wastewater data can give a good estimation of the cohort study prevalence.

For this prediction, additional parameters such as air or water temperature could also be used. This is because the prediction is done at the city level, where these parameters cannot be aggregated at federal state level (cf. "Wastewater"). We did the prediction for the cities of Koblenz, Kaiserslautern, and Mainz since for those, the investigated wastewater parameters are available. The prevalence was predicted for each city individually and the prevalence estimator for Rhineland-Palatinate was computed by combining the individual prevalences. As described in "Statistics", random forests with leave-one-out cross-validation were fitted.

Figure 5 shows the results. The left panel compares the true prevalence with the random-forest-predicted prevalence. The cohort study prevalence cannot be estimated well from the wastewater data. In particular, the predicted prevalence appears to be quite constant and the peak in late February is not detected. In the right panel, the feature importances are depicted. The most relevant parameters are the genecopies per ml at the same measurement, the genecopies per ml two measurements before, and the absolute number of genecopies at the last measurement. Apart from the Covid-19 genecopies in wastewater, the most relevant parameter seems to be the air temperature and the flow.

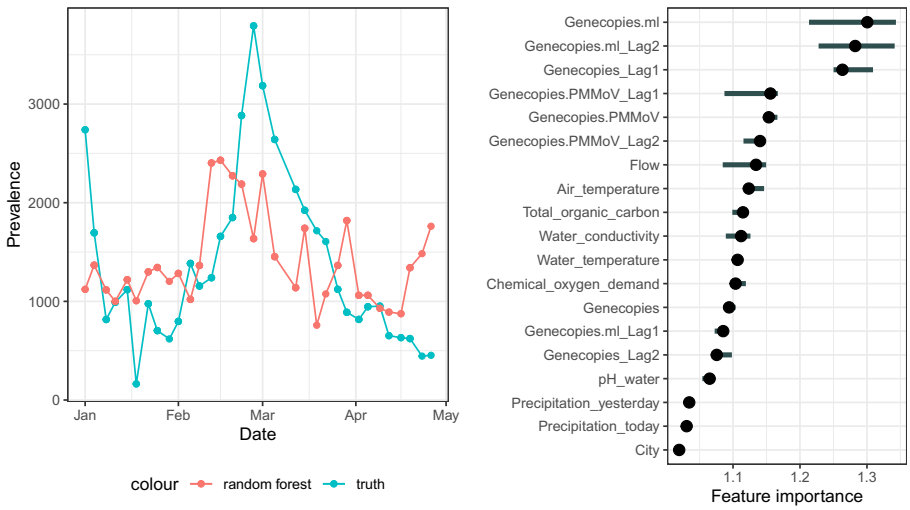

Figure 5. Prevalence prediction and corresponding feature importance derived by a random forest. The wastewater data do not allow a good prediction of the cohort study prevalence. The term 'lag' defines the measurement 'lag' before the current measurement.

Discussion

In this paper, we described the different Covid-19 surveillance techniques that were applied in Rhineland-Palatinate, Germany. We investigated how the sources wastewater data, number of hospitalizations, and prevalence of a cohort study correlate and how they can be used to track the progress of the pandemic. We demonstrated that trends in wastewater are present for all three common normalization techniques (genecopies absolute, genecopies per ml, genecopies per references virus) and that in a qualitative manner, pandemic waves can be detected in all three data sources. Furthermore, it was shown that in particular the cohort study prevalence may be well suited to build a data-driven prediction model for the future number of Covid-19-related hospitalizations. While, to the best of our knowledge, this is the first publication comparing all these three data sources, it confirms the finding that qualitative pandemic trends can be detected in wastewater[3,4,6].

There are some limitations that have to be noted to this observation. First, there is typically a time delay in collecting wastewater samples, analyzing the genecopies, and collecting the results of the cohort study patients. As a result, the data is not usually available immediately. To develop a valid prediction tool, the development of a fast data collection process would be essential. Second, the Omicron variant was the dominating variant during our entire observation period. During a pandemic, there are different variants that are dominating for a certain period of time. These variants may produce more or less genecopies in wastewater and lead to more or less hospitalizations. Consequently, a regular variant sequence analysis would be a necessary complement of an established pandemic observation tool built on wastewater. Of note, from February on, the dominated subvariant changed from BQ.1 to XBB.1. This suggests that the presented methodology is at least stable with regard to this subvariants. Third, the observed trends are mainly detectable due to the applied smoothing technique. These techniques are not valid at the boundaries of the observed time period and thus not well-suited to do extrapolation and predictions. A mathematical model that uses biological information might be a valid alternative to predict the future progress of the pandemic.

When setting up such a framework, different challenges have to be taken into account. For the wastewater data, wastewater has to be collected in sewage plants and transported to and analyzed by a laboratory. For the cohort study, one has to build a medical team, select and recruit the participants, send the tests, and establish a feedback loop. The resulting data has to be analyzed by statisticians. All these steps incur costs of time, money, and resources. Regarding the financial costs, the cohort study is more than twice as expensive as the wastewater data collection.

There are three possible extraction techniques for taking wastewater samples. One can extract time-proportionally, flow-proportionally, or volume-proportionally[32]. To obtain the most representative sample, it is recommended to apply volume- or flow-proportional extraction[7]. However, in our sample, all three extraction techniques were applied. To establish a regular tool for wastewater surveillance, it may be sound to apply the same extraction technique among different sewage plants and to avoid time-proportional extraction.

We aggregated the available data to the level of the federal state Rhineland-Palatinate, since the hopsitalization data was only available on this level. However, the cohort study was performed only on the five largest cities, i.e. only in urban areas and only on adults. Although these cities represent 713,000 inhabitants, i.e. 17% of the total population, and are geographically spread around the state, there may be structural differences in the infection characteristics that lead to a bias in the measured prevalence.

The wastewater treatment plants cover a size of 15,300 to 250,000 connected inhabitants, i.e., they are chosen from both rural and urban areas. In total 1,347,318 inhabitants are connected to the sewage plants, corresponding to a proportion of 33% of the total population of Rhineland-Palatinate. However, it is also not proven that these plants are perfectly representative for the federal state.

While we treat the hospitalization data as the gold standard for the burden of disease in the population, the hospitalization data only measures hospitalized people *with* a positive test and not necessarily patients hospitalized *because of* Covid-19. It is likely, that other (respiratory) diseases such as influenza led to an overproportional rate of hospital admissions and since we measure the absolute number of positive patients, we would then see higher levels despite potentially stable disease rates.

When analyzing the information in the wastewater data by introducing other parameters than genecopies, we focused on prediction on sewage plant level and aggregated the predictions to one value for Rhineland-Palatinate afterwards (cf. "Finding 4: Wastewater data alone does not allow quantitative prediction of the cohort study prevalence") This was done since parameters as the air temperature or the pH value in water can hardly be summarized to one joined value among different cities. We observed a quite volatile cohort study prevalence on sewage plant level. This complicates the estimation of the prevalence and makes valid assertions on the benefit of wastewater data more difficult.

While in our data, an exact prediction of the cohort study prevalence by wastewater data without incorporating biological models was not possible, this does not imply that collecting wastewater data is senseless. Wastewater data may not be used for a quantitative prediction of the pandemic development but it may serve as an qualitative predictor. This means that an increase of the viral load in wastewater indicates a worsening of the pandemic situation. Furthermore, wastewater data can be gene-sequenced in contrast to a cohort study that measures positive Covid-19 tests. This implies that wastewater systems can also be used to test new variants or other viruses. There already exist approaches to influenza virus[33,34] and reviews on different viruses, including among other Hepatitis A viruses or noroviruses[35,36]. Since wastewater is easy and inexpensive to collect, it may develop to a more and more important tool to track the spread of different viruses in the population. The investigation of wastewater surveillance for further diseases is an attractive topic for future research.

Data availability

The cohort study data are presented online under https://www.unimedizin-mainz.de/SentiSurv-RLP/dashboard/index.html. The number of N1 and N2 genecopies per ml in wastewater is publicly available online under https://lua.rlp.de/unsere-themen/humanmedizin/daten-zu-atemwegserkrankungen/abwassermonitoring. A slightly differently counted version of the number of hospitalizations than used for this paper is publicly available online under https://www.healthcare-datenplattform.de/dataset/hospitalisierung. The German weather data are publicly available online under https://www.dwd.de/DE/leistungen/cdc/cdc_ueberblick-klimadaten.html.

Received: 8 August 2023; Accepted: 29 April 2024
Published online: 03 May 2024

References

1. Feng, S. *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 ES &T Water* **1**, 1955–1965. https://doi.org/10.1021/acsestwater.1c00160 (2021).
2. Diamond, M. B. *et al.* Wastewater surveillance of pathogens can inform public health responses. *Nat. Med.* **28**, 1992–1995. https://doi.org/10.1038/s41591-022-01940-x (2022).
3. Duvallet, C. *et al.* Nationwide trends in COVID-19 cases and SARS-CoV-2 RNA wastewater concentrations in the united states. *ACS Es &t Water* **2**, 1899–1909. https://doi.org/10.1021/acsestwater.1c00434 (2022).
4. Hewitt, J. *et al.* Sensitivity of wastewater-based epidemiology for detection of SARS-CoV-2 RNA in a low prevalence setting. *Water Res.* **211**, 118032. https://doi.org/10.1016/j.watres.2021.118032 (2022).
5. Langeveld, J. *et al.* Normalisation of SARS-CoV-2 concentrations in wastewater: The use of flow, electrical conductivity and crAssphage. *Sci. Total Environ.* **865**, 161196. https://doi.org/10.1016/j.scitotenv.2022.161196 (2023).
6. Cariti, F. *et al.* Wastewater reveals the spatiotemporal spread of SARS-CoV-2 in the canton of ticino (switzerland) during the onset of the COVID-19 pandemic. *Acs Es &T Water* **2**, 2194–2200. https://doi.org/10.1021/acsestwater.2c00082 (2022).
7. Wade, M. J. *et al.* Understanding and managing uncertainty and variability for wastewater monitoring beyond the pandemic: Lessons learned from the united kingdom national COVID-19 surveillance programmes. *J. Hazard. Mater.* **424**, 127456. https://doi.org/10.1016/j.jhazmat.2021.127456 (2022).
8. Robert Koch-Institut. SARS-CoV-2-PCR-testungen in deutschland. (2023). https://doi.org/10.5281/zenodo.7646187.
9. Ministerium für Wissenschaft und Gesundheit Rheinland-Pfalz. Gesundheitsminister Clemens Hoch: LUA veröffentlicht Messdaten zum Corona-Abwassermonitoring Rheinland-Pfalz. (2022). https://mwg.rlp.de/service/pressemitteilungen/detail/gesundheitsminister-clemens-hoch-lua-veroeffentlicht-messdaten-zum-corona-abwassermonitoring-rheinland-pfalz.
10. Robert Koch Insitut. Wochenberichte zu COVID-19. (2021–2023). https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Situationsberichte/Wochenbericht/Wochenberichte_Tab.html?nn=13490888.
11. Bundesministerium für Gesundheit. Hospitalisierungsinzidenz. (2021). https://www.bundesgesundheitsministerium.de/coronavirus/hospitalisierungsinzidenz.
12. Bundesministerium für Gesundheit. Nationale Teststrategie SARS-CoV-2. (2022). https://www.bundesgesundheitsministerium.de/fileadmin/Dateien/3_Downloads/C/Coronavirus/Teststrategie/NationaleTeststrategie_Schaubild.pdf.
13. Wälde, K. How to remove the testing bias in CoV-2 statistics. *medRxiv* 2020-10 (2020). https://doi.org/10.1101/2020.10.14.20212431.
14. Boehm, A. B., Wolfe, M. K., White, B., Hughes, B. & Duong, D. Divergence of wastewater SARS-CoV-2 and reported laboratory-confirmed COVID-19 incident case data coincident with wide-spread availability of at-home COVID-19 antigen tests. *PeerJ* **11**, e15631. https://doi.org/10.7717/peerj.15631 (2023).
15. Maal-Bared, R. *et al.* Does normalization of SARS-CoV-2 concentrations by Pepper Mild Mottle Virus improve correlations and lead time between wastewater surveillance and clinical data in Alberta (Canada): Comparing twelve SARS-CoV-2 normalization approaches. *Sci. Total Environ.* **856**, 158964. https://doi.org/10.1016/j.scitotenv.2022.158964 (2023).
16. Ciannella, S., González-Fernández, C. & Gomez-Pastora, J. Recent progress on wastewater-based epidemiology for COVID-19 surveillance: A systematic review of analytical procedures and epidemiological modeling. *Sci. Total Environ.* **878**, 162953. https://doi.org/10.1016/j.scitotenv.2023.162953 (2023).
17. Tang, L. *et al.* Exploration on wastewater-based epidemiology of SARS-CoV-2: Mimic relative quantification with endogenous biomarkers as internal reference. *Heliyon* **9**, (2023). https://doi.org/10.1016/j.heliyon.2023.e15705.
18. Nourbakhsh, S. *et al.* A wastewater-based epidemic model for SARS-CoV-2 with application to three Canadian cities. *Epidemics* **39**, 100560. https://doi.org/10.1016/j.epidem.2022.100560 (2022).
19. Robert Koch Institute. Systematic surveillance for SARS-CoV-2 in wastewater. (2023). https://www.rki.de/EN/Content/Institute/DepartmentsUnits/InfDiseaseEpidem/Div32/WastewaterSurveillance/WastewaterSurveillance.html
20. Kitajima, M., Sassi, H. P. & Torrey, J. R. Pepper mild mottle virus as a water quality indicator. *NPJ Clean Water* **1**, (2018). https://doi.org/10.1038/s41545-018-0019-5.
21. German Meteorological Service. Klimadaten zum direkten Download. (2023). https://www.dwd.de/DE/leistungen/cdc/cdc_ueberblick-klimadaten.html
22. Rheinland-Pfalz, L. SARS-CoV-2-Abwassermonitoring für Rheinland-Pfalz. (2023). https://lua.rlp.de/unsere-themen/humanmedizin/daten-zu-atemwegserkrankungen/abwassermonitoring
23. Wild, P. Vorstellung von SentiSurv RLP. (2023). https://www.unimedizin-mainz.de/sentisurv/ueber-sentisurv-rlp/vorstellung-von-sentisurv-rlp.html
24. Wild, P. Dashboard SentiSurv RLP. (2023). https://www.unimedizin-mainz.de/SentiSurv-RLP/dashboard/index.html
25. Van den Bossche, J. *et al.* Geopandas/geopandas: v0.14.0. (2023). https://doi.org/10.5281/zenodo.8348034.
26. Wilson, E. B. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.* **22**, 209–212. https://doi.org/10.1080/01621459.1927.10502953 (1927).
27. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32. https://doi.org/10.1023/a:1010933404324 (2001).
28. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2023).
29. Wickham, H. *et al.* Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019). https://doi.org/10.21105/joss.01686.
30. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2**, 18–22 (2002).
31. Molnar, C., Bischl, B. & Casalicchio, G. Iml: An r package for interpretable machine learning. *J. Open Source Softw.* **3**, 786 (2018). https://doi.org/10.21105/joss.00786.
32. Sandoval, S., Bertrand-Krajewski, J.-L., Caradot, N., Hofer, T. & Gruber, G. Performance and uncertainties of TSS stormwater sampling strategies from online time series. *Water Sci. Technol.* **78**, 1407–1416. https://doi.org/10.2166/wst.2018.415 (2018).
33. Mercier, E. *et al.* Municipal and neighbourhood level wastewater surveillance and subtyping of an influenza virus outbreak. *Sci. Rep.* **12**, (2022). https://doi.org/10.1038/s41598-022-20076-z.

34. Wolfe, M. K. *et al.* Wastewater-based detection of two influenza outbreaks. *Environ. Sci. Technol. Lett.* **9**, 687–692. https://doi.org/10.1021/acs.estlett.2c00350 (2022).
35. Xagoraraki, I. & O'Brien, E. Wastewater-based epidemiology for early detection of viral outbreaks. In *Women in Water Quality* 75–97 (Springer International Publishing, 2019). https://doi.org/10.1007/978-3-030-17819-2_5.
36. McCall, C., Wu, H., Miyani, B. & Xagoraraki, I. Identification of multiple potential viral diseases in a large urban center using wastewater surveillance. *Water Res.* **184**, 116160. https://doi.org/10.1016/j.watres.2020.116160 (2020).

Acknowledgements

We would like to thank Daniel Stich, Wolfgang Lehnen, and Markus Hies for the access to all necessary data sources. Furthermore, we would like to thank Prof. Philipp Wild and his lab for the processing of the cohort study data.

Author contributions

M.P.: Conceptualization, methodology, software, formal analysis, data curation, writing-original draft, visualization. K.-H.K.: Supervision, writing-review and editing, funding acquisition. J.M.: Conceptualization, data curation, writing-review and editing, formal analysis. J.M.: Data curation, writing-review and editing, formal analysis. J.W.: Software, data curation, writing-review and editing. N.L.: Conceptualization, formal analysis, data curation, writing-original draft, project administration.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work has been supported by the Ministry for Science and Health of Rhineland-Palatinate.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.