Analysis of SARS-CoV-2 in wastewater for prevalence estimation and investigating clinical diagnostic test biases

Mattia Mattei [a,*], Rosa M. Pintó [b], Susana Guix [b], Albert Bosch [b], Alex Arenas [a,c,**]

[a] Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, 43007 Tarragona, Spain
[b] Enteric Virus Laboratory, School of Biology, University of Barcelona, 08028, Barcelona, Spain
[c] Pacific Northwest National Laboratory, 902 Battelle Blvd, Richland, WA, 99354, USA

ARTICLE INFO

ABSTRACT

Here we analyze SARS-CoV-2 genome copies in Catalonia's wastewater during the Omicron peak and develop a mathematical model to estimate the number of infections and the temporal relationship between reported and unreported cases. 1-liter samples from 16 wastewater treatment plants were collected and used in a compartmental epidemiological model. The average correlation between genome copies and reported cases was 0.85, with an average delay of 8.8 days. The model estimated that 53% of the population was infected, compared to the 19% reported cases. The under-reporting was highest in November and December 2021. The maximum genome copies shed in feces by an infected individual was estimated to range from 1.4×10^8 gc/g to 4.4×10^8 gc/g. Our framework demonstrates the potential of wastewater data as a leading indicator for daily new infections, particularly in contexts with low detection rates. It also serves as a complementary tool for prevalence estimation and offers a general approach for integrating wastewater data into compartmental models.

1. Introduction

The emergence of the SARS-CoV-2 coronavirus in 2019 has resulted in a global pandemic, which, as October 2022, has led to over 600 million infections and 6 million deaths worldwide. Epidemiological data has played a crucial role in monitoring the spread of the epidemic, with clinical testing via reverse transcription quantitative polymerase chain reaction (RT-qPCR) on nasopharyngeal swabs being the primary method. However, limitations and biases exist in any epidemiological indicator, particularly in the case of PCR testing, which relies on voluntary participation and may only capture individuals more likely to be infected. Given the high number of asymptomatic and subclinical infections, this biased testing process can significantly impact estimations. Although hospitalizations and deaths are less susceptible to bias, they may not be ideal for real-time forecasting due to their lagged estimations.

Wastewater-based epidemiology (WBE), i.e. the surveillance of epidemic spreading through the analysis of virus concentration in wastewater plants, is therefore presenting itself as a potential complementary tool to clinical testing, and it is gaining more and more attention among the mathematical modelers. The concept of WBE centers around the knowledge that SARS-CoV-2 RNA can be detected in stool samples

excreted by human bodies (Wolfel et al., 2020; Wang et al., 2020; Carcereny et al., 2022a), and then shed in the sewage system. Therefore, daily sampling of SARS-CoV 2 RNA in wastewater would provide information similar to that from daily random testing of thousands of individuals in a community (Larsen and Wigginton, 2020), but not distinguishing between symptomatic, asymptomatic or presymptomatic people as long as they develop viral RNA in their feces. The interest of the WBE relays on two main aspects: wastewater data can potentially account the unreported cases, and they can also represent an estimate in advanced over time respect to diagnostic tests. Consequently, WBE is envisaged to become the most important non-invasive diagnostic tool of the epidemics in a population.

The concept of wastewater epidemiology has frequently been referred to as a "leading indicator" of reported cases (Olesen et al., 2021), although there is often a time delay between the two measures (Peccia et al., 2020; D'Aoust et al., 2021; Morvan et al., 2022; Wu et al., 2020b). The extent of the lead time provided by sewage data varies significantly in the literature, ranging from a few days to up to two weeks (Krivoakova et al., 2021). The duration of the time delay between wastewater estimates and reported cases is influenced by various factors, including the characteristics of the health system such as the

---

* Corresponding author.
** Corresponding author at: Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, 43007 Tarragona, Spain.
E-mail addresses: mattia.mattei@urv.cat (M. Mattei), alexandre.arenas@urv.cat (A. Arenas).

availability and distribution of diagnostic tests, and the time required to obtain test results (Olesen et al., 2021). However, it is important to note that wastewater estimates are subject to considerable uncertainty, which is attributable to several factors. Firstly, our understanding of the shedding process is limited. As detailed in Section 2.2, the quantity of SARS-CoV-2 shed in feces and its temporal profile exhibit considerable inter-individual variability, and clinical studies have reported a wide range of results. Moreover, it remains unclear whether the onset of viral shedding in feces precedes or coincides with the onset of symptoms, given that most existing clinical studies have been conducted in hospitalized patients.

Secondly, the virus within the sewage system is subject to various "random" factors, such as dilution with the daily water flow, temperature, possible interactions with chemical agents or other substances, and environmental factors like rain. The features of the sewage systems, such as the travel time from households to treatment plants, can also have a significant impact on measurements. Lastly, the experimental process for extracting data on genetic copy concentration from sewage is not without challenges, and measurement errors should always be taken into consideration.

Upon considering the previous explanations, a fundamental question arises as to how to mathematically quantify the biases between the genome copies concentration in wastewater, the reported cases, and the actual incidence in any given area. To this end, we conducted an analysis of data pertaining to the absolute concentrations of the SARS-CoV-2 gene biomarker N1 in weekly wastewater samples collected from 16 wastewater treatment plants (WWTP) in Catalonia, Spain, during the period spanning October 2021 to March 2022. The data was sourced from the Catalan Surveillance Network of SARS-CoV-2 in Sewage (https://sarsaigua.icra.cat/). Initially, we examined the time delay and statistical correlation between wastewater data and reported cases at each WWTP. The data pertaining to reported cases was obtained from the official website of the Catalonia government (Generalitat de Catalunya https://analisi.transparenciacatalunya.cat/browse{q}=covid&sortBy=relevance). Subsequently, we proposed a model that incorporates a time-varying rate of unreported cases to explain the observed delays and, in general, the heterogeneity of outcomes reported in the literature on the subject.

The principal outcomes of our study are twofold: firstly, our analysis of wastewater data in the Catalonia region reveals a markedly high correlation with reported cases, with a mean Pearson correlation of 0.85, and an average 9-day advance in anticipating trends in reported cases, but with variability ranging from 0 to 20 days. Secondly, the proposed model enables us to successfully link wastewater data with temporal dynamics of the reported cases during the same period, and provides estimates of the actual prevalence of infection and parameters of interest in the context of wastewater-based epidemiology.

2. Methods

2.1. Wastewater sampling

The study involved the weekly collection of 1-liter composite samples of influent wastewater from 16 wastewater treatment plants (WWTPs) (Table 1) in the region of Catalonia, Spain. The samples were collected over a period of six months, from October 2021 to March 2022. The WWTPs selected covered a population of 2,514,618 inhabitants, which represents around 31% of the total Catalan population. This information was obtained summing up the population of all the municipalities served by the plants considered. The samples were transported in a portable icebox at a temperature range of 0°–4 °C and were analyzed the day after concentration. The wastewater samples were collected within the framework of the Catalan Surveillance Network of SARS-CoV-2 in Sewage (https://sarsaigua.icra.cat/; https://doi.org/10.5281/zenodo.4147073).

2.2. Wastewater concentration

Wastewater samples were concentrated by using the aluminum hydroxide adsorption–precipitation method, as described in previous studies (Randazzo et al., 2020; Wallis and Melnick, 1967). Two hundred milliliters of wastewater were concentrated to a final volume of 1–5 milliliters of phosphate-buffered saline (PBS). To ensure the accuracy of the concentration process, 1.5×10^6 TCID50 units of the attenuated PUR46-MAD strain of the Transmissible Gastroenteritis Enteric Virus (TGEV) (Moreno et al., 2008) were seeded into each sample prior to the concentration step.

2.3. Nucleic acid extraction

To extract nucleic acids, 300 µl of the concentrated samples were used and the Maxwell® RSC PureFood GMO and Authentication Kit (Promega) was employed following the manufacturer's instructions. In each extraction run, a PBS negative control was included. To determine virus recovery, a previously described RT-qPCR assay for quantification of the Transmissible Gastroenteritis Enteric Virus (TGEV) was used (Vemulapalli et al., 2009). Samples with virus recovery ≥1% were deemed acceptable, following criteria established in ISO 15216-1:2017 & ISO 15216-2:2020 for the analysis of norovirus and hepatitis A virus in food and water samples. Recoveries from the present data set were from 1% to 100% with an average of 36.68% ± 21.60%.

2.4. RT-qPCR assays

Quantification of SARS-CoV-2 RNA in sewage samples was based on the N1 assay (US-CDC 2020), which targets a fragment of the nucleocapsid gene. We employed the PrimeScript™One Step RT-PCR Kit (Takara Bio, USA) and a CFX96 BioRad instrument.

RT-qPCR data analysis and interpretation were performed as previously described (Carcereny et al., 2021, 2022b). RT-qPCR analysis included the analysis of duplicate wells containing undiluted RNA and duplicate wells containing a 10-fold dilution to monitor the presence of inhibitors. Every RT-qPCR assay included four wells corresponding to negative controls (two nuclease-free water and two negative extraction controls). Commercially available Twist Synthetic SARS-CoV-2 RNA Controls were used to prepare standard curves for genome quantitation. For each specific target, Cq values ≤ 40 were converted into gc/L using the corresponding standard curve and volumes tested. Occurrence of inhibition was estimated by comparing average viral titers obtained from duplicate wells tested on undiluted RNA with duplicate wells tested on 10-fold diluted RNA. Inhibition was ascertained when difference in average viral titers was higher than 0.5 log10, and if that occurred, viral titers were inferred from the 10-fold RNA dilution.

2.5. Convolution description of viral shedding

Convolution operations represent the most appropriate approach to mathematically model the relationships between genome copy concentration, reported cases, and actual infection prevalence. Convolution is a mathematical method that involves the combination of two functions to produce a third function that describes how one function modifies the other. The resulting function is defined as the convolution of the two input functions. In essence, the procedure involves sliding one function over the other, multiplying the overlapping portions of the two functions, and integrating the product over the entire variable range to generate a novel function that portrays the interplay between the two functions.

The virus concentration in sewage can be modeled as a function of the number of infected individuals in the serviced area and the time since they became infected, given a specific profile of the quantity of virus shed in feces over time.

Table 1

The maximum Pearson correlations and the corresponding delays between sewage data and reported cases for each WWTP. For each of the 16 wastewater plants listed in the left column we measured the Pearson correlation between genome copies concentrations data linearly interpolated and 7-days averaged daily reported cases in the corresponding served municipalities. We performed the analysis shifting back in time reported cases from 0 to 20 days. We interpreted as *delay* the shift at which we reached maximum correlation.

WWTP	Max. correlation	Delay (days)
BANYOLES	0.96	4
RUBÍ	0.96	13
MARTORELL	0.96	12
LLAGOSTA_LA	0.94	5
PRAT_DE_LLOBREGAT_EL	0.92	5
FIGUERES	0.9	20
VILAFRANCA_DEL_PENEDÈS	0.87	3
ABRERA	0.85	0
MONTCADA	0.85	20
GRANOLLERS	0.83	20
PUIGCERDÀ	0.83	5
BERGA	0.82	0
MONTORNÈS_DEL_VALLÈS	0.79	0
GIRONA	0.78	10
RIERA_DE_LA_BISBAL	0.74	20
SABADELL/RIU_SEC	0.64	4

There is a general consensus in the scientific literature on certain characteristics of the virus shedding profile: long duration, exponential decay, and peak around symptom onset. Wu et al. (2020a) reported that SARS-CoV-2 RNA can be detected in feces for a mean of 11.2 days after respiratory tract samples test negative (up to 5 weeks); Zhang et al. (2020) found a median fecal shedding duration of 22 days. Wolfel et al. (2020) observed RNA-positive stool samples for over 3 weeks without symptoms, with peak viral RNA likely occurring during the first week of symptoms. The timing of shedding onset relative to symptom onset remains debated due to lack of clinical data on exposed individuals, but Hoffmann and Alsing (2021) constrained the latter part of the shedding profile with a fast exponential decay. Miura et al. (2021) successfully tested the model proposed by Teunis et al. (2015) for norovirus shedding, to SARS-CoV-2 clinical data, which accounts for both exponential rise and decay. Many authors employed gamma or beta distributions to describe viral shedding in feces. The beta distribution is a symmetrical distribution while the gamma distribution is unimodal with a skewed shape. However, the beta distribution is defined only on [0,1] and it needs to be rescaled. In general, they both can be used as showed by Wu et al. (2020b) who found very similar results for the two descriptions. Wrapping up, we retain that a description that accurately captures the essential characteristics of the viral shedding profile is a gamma distribution, as reported in previous studies such as Huisman et al. (2022) and Fernandez-Cassi et al. (2021). Specifically, Huisman et al. (2022) used data on the incubation period and gastrointestinal shedding following symptoms onset to model the shedding profile as a gamma distribution with a mean of 6.7 days and a standard deviation of 7 days. We adopted this approach in our analysis.

Therefore, we modeled the quantity of genome copies at day t as

$$CG(t) = \bar{k} \sum_{t'=t-30}^{t'=t} \Gamma(t-t') N_I(t'),\tag{1}$$

where the number of new infections at time t', $N_I(t')$, is convoluted with the gamma distribution described above, truncated at 30 days, which tell us the quantity of genome copies per gram of feces shed at $t - t'$ days after the infection. The factor $\bar{k}$ is a scale parameter and it should take in account several aspects: the degradation, D (defined between 0 and 1), that the shed virus may undergoes in his way to the plant (this is affected by multiple factors like water temperature, dilution, chemical reactions as well as by the time the virus spends in it), the average grams of feces produced per person g, the fraction of

infected people shedding virus in feces p and the total quantity of virus shed in a gram of feces by an individual during the entire course of the infection Q. Therefore, similarly as in Ahmed et al. (2020), we can write

$$\bar{k} = Q \times p \times g \times D.\tag{2}$$

Following Chavarria-Miro et al. (2021) the quantity g can be taken equal to 380 grams per day, based on an excretion of 30 g per 5.5 kg of body weight, assuming an average weight of Spanish population of 70 kg (https://www.mscbs.gob.es/estadEstudios/sanidadDatos/). The value for the fraction of infected people shedding virus in feces p, is quite variable in the literature, ranging from the 29% of Wang et al. (2020) to the 83.3% of the patients for Zhang et al. (2020). The review on the topic by Cheung et al. (2020) suggests to consider a value equal to the 48.1%.

The value of Q is also uncertain. Several studies (Wolfel et al., 2020; Cheung et al., 2020; McMahan et al., 2021; Xu et al., 2020) agree that the maximum possible shed quantity of genome copies per milliliter of stool should be around $Q_{max} = 10^7$ gc/ml; Zhang et al. (2020) indicates one order of magnitude less ($Q_{max} = 10^{5.8}$ gc/ml) while Arts et al. (2022) proposed a value around $Q_{max} = 10^9$ gc/g. The dissipation process is also quite difficult to describe as affected by random factors and, in principle, it could change in time. McMahan et al. (2021) proposed to use an exponential decay model which considers the effects of the water temperature and of the holding time on the virus. As stated by Weidhaas et al. (2021), reported decay rates for SARS-CoV and surrogate coronaviruses in unpasteurized wastewater at 23 °C range from 0.02 to 0.143 per hour. Bivins et al. (2020) found that the time for 90% reduction (T90) specifically for viable SARS-CoV-2 in wastewater at room temperature was 1.5 days. In wastewater at 50 and 70 °C, the observed T90 values for infectious SARS-CoV-2 were decreased to 15 min. The latter, among other works (McCall et al., 2022; Phan et al., 2023), show how temperature and travel-time can largely affect the value for the decay rate D. However, we want to remark here that in our theoretical framework we will not make any assumptions about the value of the decay rate or for the other terms in Eq. (2), while the general scale factor $\bar{k}$ will be evaluated automatically from the data through the parameters estimation procedure (see Section 2.7.2).

2.6. Compartmental model with a time-varying rate of reported infections

Compartmental models, specifically ordinary differential equation (ODE) models, have been the cornerstone of infectious disease modeling for over a century. These models divide the population into different compartments based on their infectious status, such as susceptible, infected, and recovered in the classical SIR model (Kermack and McKendrick, 1927). The movements of individuals between compartments are described by transition rates, which are based on the underlying biology of the disease, as well as demographic and behavioral factors. By simulating these transitions, compartmental models can be used to predict the future course of an outbreak and to evaluate the impact of different intervention strategies (Arenas et al., 2020).

In this study, we propose a variation of the Susceptible–Infected–Recovered (SIR) model, where infected individuals are divided into those who are infected but not detected (I_N) and those who are detected and isolated (I_D). The model is described by a system of differential equations, where the transmission rate of the disease is represented by β, the recovery rate by γ, and the total population by N:

$$\frac{dS(t)}{dt} = -\frac{\beta S(t) I_N(t)}{N},\tag{3}$$

$$\frac{dI_N(t)}{dt} = \frac{\beta S(t) I_N(t)}{N} - \gamma I_N(t) - p(t) I_N(t),\tag{4}$$

$$\frac{dI_D(t)}{dt} = p(t) I_N(t) - I_D(t),\tag{5}$$

$$\frac{dR(t)}{dt} = \gamma I_N(t) + I_D(t).\tag{6}$$

Note that we have included in the model a time-dependent probability of infected individuals being detected, represented by $p(t)$, which is proportional to the ratio of the detected infections at time t:

$$p(t) = p_0 + (1 - p_0)\left(\frac{I_D(t)}{I_N(t) + I_D(t)}\right), \tag{7}$$

where, in case of zero detection, $p(t) = p_0$. This equation consists of a constant part and a time-dependent one: at each time step there is a constant percentage p_0 of infected that decide to get tested *unconditionally* while the rest is more sensible to the available information about the actual state of the epidemics.

This probability is influenced by factors such as changes in testing availability, policy, and implementation of Non-Pharmaceutical-Interventions (NPIs), as well as the general perception of the population about the ongoing epidemic. We assume that infected individuals, once detected, are automatically removed from the infected but not detected compartment. Moreover, we argue that $p(t)$ is also fundamentally related to the general perception that the population has about the on-going epidemic, especially when the testing process is subministered on voluntary basis: people can be more or less willing to be tested according to their risk awareness or according to costs/benefits considerations, which clearly depend on the state of the epidemic, or better, on its *perceived* state; all the information that people have about the epidemic are enclosed in the daily reported cases. Given this and taking inspiration by several works which tried to model risk perception (Poletti et al., 2011, Steinegger et al., 2020).

Our idea is to validate this model using both wastewater data and reported cases information. The former can be generated at each time step according to Eq. (1), with daily new infections estimated by the system of equations above.

We argue that our model has the potential to provide insights into parameters of interest, such as $\bar{k}$ and $p(t)$, which can justify the spectrum of delays observed between wastewater data and reported cases. This theoretical framework can provide valuable information about the dynamics of infectious diseases and can inform public health policy and decision-making.

2.7. Statistical methods

2.7.1. Pearson Correlation Coefficient

The Pearson Correlation Coefficient (r) is a statistical measure commonly used to assess the strength of the linear relationship between two variables. The formula for the Pearson correlation coefficient is based on the covariance between two variables, normalized by their respective standard deviations. When applied to a sample, the Pearson coefficient is usually represented as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{8}$$

where n is the sample size, x_i and y_i are the individual values of the two variables, $\bar{x}$ and $\bar{y}$ are their respective means. The value of r ranges from -1 to $+1$, where a value of -1 indicates a perfect negative correlation, a value of $+1$ indicates a perfect positive correlation, and a value of 0 indicates no correlation. The coefficient assumes that the relationship between the variables is linear and that the data is normally distributed and it is also quite sensitive to outliers.

In this work we used the coefficient to assess correlation between wastewater data and *shifted* daily reported cases time-series. The aim was to compute the value of the shift (in terms of days) for which we obtained maximum correlation and, therefore, to quantify the delay between the two curves. In this context and with no modeling intentions, the assumptions for linearity and normality are deemed acceptable.

2.7.2. Approximate Bayesian computation

Bayesian inference allows to make probabilistic statements about unknown parameters of a model given observed data. The central idea of Bayesian inference is to update our prior beliefs about the parameters using the observed data and obtain a posterior distribution that reflects our updated knowledge about the parameters. Bayes' rule provides a mathematical formula for computing the posterior distribution of the parameters given the observed data.

However, defining and computing the likelihood function can be challenging when the model is complex or when the observed data are high-dimensional. Approximate Bayesian Computation (ABC) is a simulation-based approach that circumvents the need to compute the likelihood function directly. Instead, ABC generates synthetic data sets from proposed parameter values using a forward model and compares them to the observed data using a distance metric. The parameter values that produce synthetic data sets that are close to the observed data are retained, representing a sample from the approximate posterior distribution. The ABC algorithm can be formulated as follows:

- Draw a proposed parameter value θ from the prior distribution $p(\theta)$;
- Simulate synthetic data y^* from the forward model using θ;
- Compute the distance between the observed data y and the synthetic data y^* using a distance metric $d(y, y^*)$;
- If $d(y, y^*)$ is smaller than a specified tolerance ϵ, accept θ and add it to the retained sample. Otherwise, reject θ and return to step 1;
- Repeat steps 1–4 until a sufficient number of parameter values have been retained.

We calibrated our model using PyMC[1] package in python 3.7 which used Markov Chain Monte Carlo (MCMC) methods. The disadvantage of the classical ABC rejection sampler is that the acceptance rate is low when the prior distribution is very different from the posterior distribution (Toni et al., 2009). To avoid this problem, an ABC method based on Markov Chain Monte Carlo was introduced (Marjoram et al., 2003) to improve the efficiency of sampling from the approximate posterior distribution. The ABC-MCMC algorithm generates a sequence of parameter values that converge to the approximate posterior distribution.

3. Results

3.1. Statistical description

We calculated the Pearson correlation between the weekly samples of number of genome copies in each wastewater treatment plant (linearly interpolated) and the 7-days averaged number of daily reported COVID-19 cases for each specific plant. We also performed a smoothing procedure of the wastewater time-series through a Savitzky–Golay filter in order to filterer out the outliers. The reported cases were shifted back from 0 to 20 days to quantify the delay between sewage data and reported cases. We analyzed the period between October 2021 and March 2022, during which the Omicron variant was spreading rapidly in Catalonia and other parts of the world. The results, summarized in Table 1 and Fig. 1, showed an average correlation of 0.85 ± 0.09 (0.64–0.96) and an average delay of 8.8 ± 7.4 (0–20) days across the 16 WWTPs.

These findings suggest that wastewater data can broadly capture the current trend of the epidemic, or at least to the extent that reported cases do. Furthermore, they seem to anticipate voluntary testing by a relevant quantity of days, more than reported in other studies. The observed delay was highly heterogeneous, ranging from 0 to 20 days. Fig. 2 compares the genome copies per liter averaged on all the 16 WWTPs versus the cases reported for the entire Catalonia, both
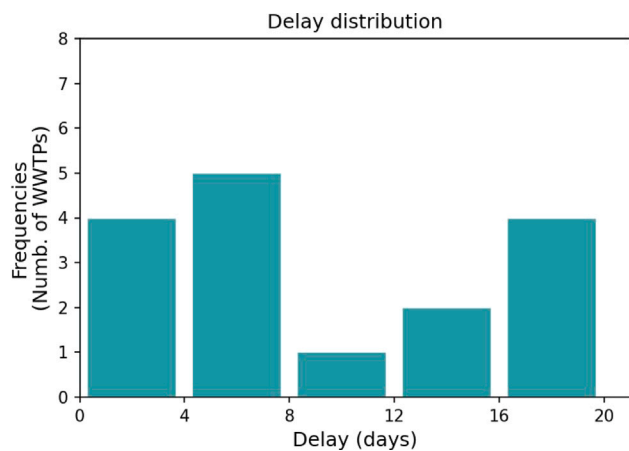

Fig. 1. Distribution of the delays between wastewater data and reported cases curve. Histogram showing the distribution of the delays between weekly wastewater data (linearly interpolated) and 7-days averaged daily reported cases curve across the 16 WWTPs considered. With *delay* we refer to the amount of days by which daily reported cases needed to be shifted back in time in order to obtain maximum Pearson correlation. This procedure results in a average correlation of 0.85 ± 0.09 (0.64–0.96) and an average delay of 8.8 ± 7.4 days (0–20).
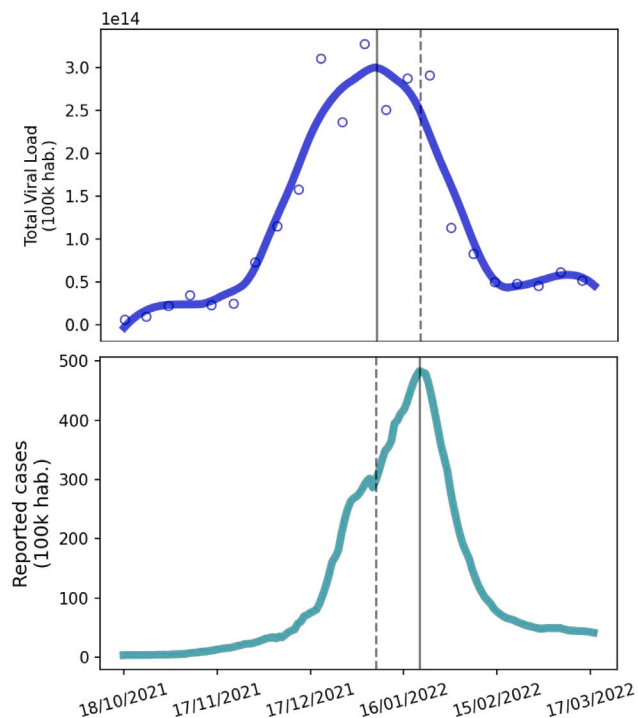
Fig. 2. Total viral load versus reported cases for Catalonia. In the top panel is displayed the weekly-sampled total viral load averaged across the 16 WWTPs (dots) and its interpolated and smoothed representation (solid line). The bottom panel shows the time-series for 7-days averaged reported cases in Catalonia (solid line). Both measures were calibrated to 100.000 inhabitants. The two vertical lines indicate the temporal distance between the peaks of the two curves.

normalized according to a population of 100,000 inhabitants. In this global perspective, the time-shift between the two curves is 14 days, with a correlation of 0.95. The model will provide plausible arguments to realistically explain a delay higher than expected, considering the available information about incubation period, fecal shedding, infection duration, and in general, to justify the wide range of observations in the literature.

3.2. Model calibration and validation

The model has been calibrated with real data of reported cases and genome copies concentration, averaged across all the 16 plants and normalized to 100,000 inhabitants, using Approximate Bayesian Computation (ABC) (Marjoram et al., 2003). For a total time period of 152 days, we trained the model with the first 100 days and then we validated it for the remaining ones. The procedure converged yielding the posterior distributions for parameters β, p_0 and $\bar{k}$ displayed in Fig. 3. The parameter γ has been chosen equal to 10 days^{-1}. All the details about the parameters can be found in Table 2.

Afterwards, we ran the model using the average parameter values obtained from the calibration process. The resulting epidemiological scenario is presented in Fig. 4, which shows the proportions of susceptible individuals (S), undetected infected individuals (I_N), detected infected individuals (I_D), and recovered individuals (R).

According to the model, approximately 53% of the population under study was infected during the period analyzed. Fig. 5 presents a comparison between the confirmed cases data, wastewater data, and the model's predictions in the left and right panels, respectively. The R^2 statistics for reported cases and genome copies are 0.94 and 0.8, respectively.

The figures reveal a significant agreement in both qualitative and quantitative terms for all stages of the epidemic wave, particularly in the case of reported cases. Sewage data, which are subject to notable fluctuations, show a lesser degree of agreement.

3.3. Detection rate and under-reporting

Our study indicates that the actual number of infections during the period of October 2021 to February 2022 in the analyzed areas of Catalonia was approximately three times higher than the reported cases. However, during November to December 2021, this ratio reached values up to ten (left panel of Fig. 6). As a result, the detection rate, which is represented by $p(t)$ in the equations, appears to be a monotonically increasing function over time (central panel in Fig. 6).

The model predicted that the daily genome copies would peak approximately five days before the simulated detected infections, which is consistent with some findings in previous literature (Peccia et al., 2020; Morvan et al., 2022). This suggests that the observed delays can be attributed to two factors: (i) fluctuations and noise in sewage data, and (ii) the value of the parameter p_0, which is related to the initial value and variability of the detection rate over time. The delay between simulated genome copies and reported cases was observed to be a monotonically decreasing function with p_0 in the equations, with values between 0.001 and 0.01 resulting in a wide range of delays (2–16 days), consistent with numerous available datasets (right panel of Fig. 6).

3.4. Maximum quantity of genome copies shed in feces by an individual

Our theoretical framework provides an estimate of the parameter $\bar{k}$ (see Section 2.5) that relates the viral load introduced into the system to that being measured. Using the deterministic Eq. (2), we estimated Q_{max}, which represents the maximum quantity of genome copies shed in a gram of feces by an individual during the course of infection. This quantity is of interest in the field of Wastewater-Based Epidemiology (WBE) applied to SARS-CoV-2 but has large fluctuations in estimations available in the literature.

Fig. 7 shows a colormap indicating the values of Q_{max} inferred from Eq. (2) using the mean value of $\bar{k}$ in the posterior distribution yielded by the model calibration procedure. We considered the possible range of values for the dissipation factor D (0.86–0.98) and the fraction of people shedding virus in feces p (0.29–0.83) according to the considerations made in Section 2.5. The results indicate a value of Q_{max} between 1.4×10^8 gc/g and 4.4×10^8 gc/g.

Table 2

Parameters of the model. The parameters were estimated using Approximate Bayesian Computation (ABC). Only the recovery rate γ was assumed to be equal to 10 days^{-1}.

Symbol	Description	Estimates	Assignment
β	Infectivity	0.225 (97% CI: 0.223 - 0.227)	Calibrated
γ	Recovery rate	10 days^{-1}	Assumed
p_0	Initial testing rate	0.004 (97% CI: 0.002 - 0.006)	Calibrated
$\bar{k}$	Scale factor for shedding process	3.16×10^{11} (97% CI: 3.14×10^{11} - 3.18×10^{11})	Calibrated
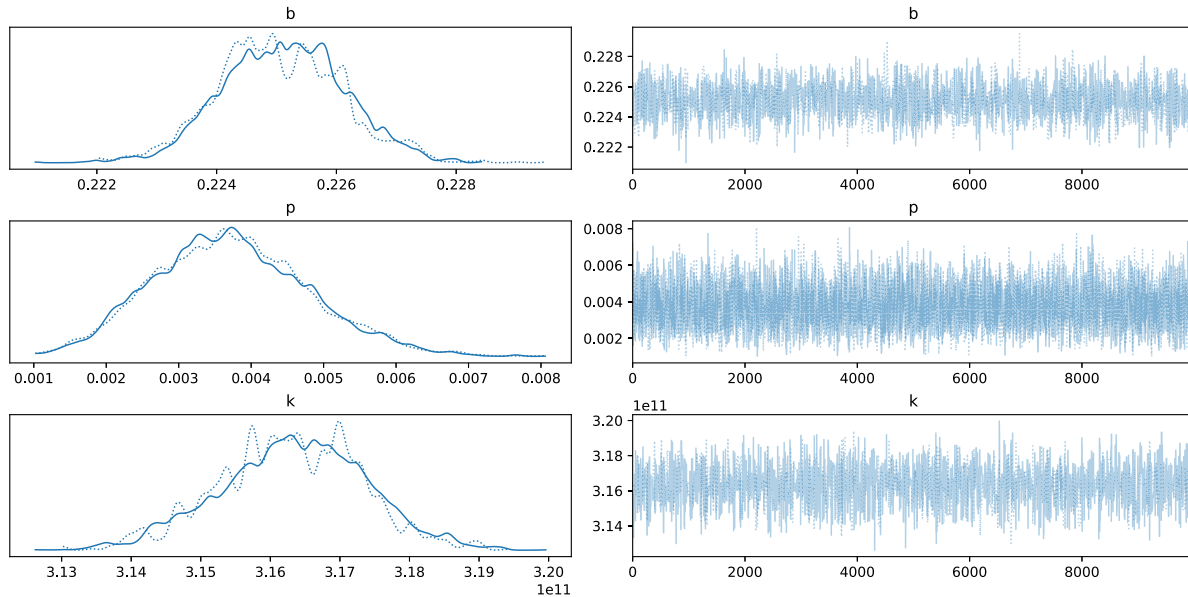
Fig. 3. Posterior distributions for parameters β, p_0 and $\bar{k}$. The figure shows the results of the Approximate Bayesian Computation (ABC) procedure for the estimation of the parameters β, p_0 and $\bar{k}$, respectively indicated as b, p and k. The plots on the left side are the posterior distributions whereas in the right side are showed all the values sampled during the process for all three parameters, which took 10,000 steps (x-asis).
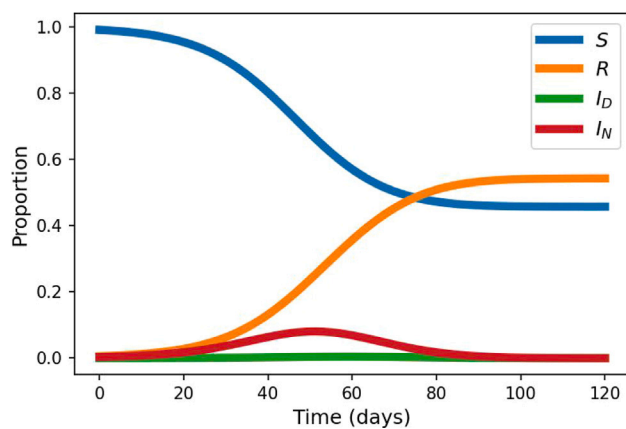
Fig. 4. Proportions of the population being susceptible (**S**), **infected not detected** (**I_N**), **infected detected** (**I_D**) and **recovered** (**R**). The figure displays the temporal evolution of the simulated epidemic according to the model. The curves indicates the proportion of susceptible (blue line), infected not detected (red line), infected detected (green line) and recovered (orange line) compartments. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4. Discussion

The aim of this study was to assess the potential of using wastewater-based epidemiology (WBE) to anticipate reported cases and estimate the actual prevalence of SARS-CoV-2 infections. We have used real data from Catalonia. The results showed that wastewater data displayed a high correlation with reported cases, indicating that WBE can capture the current trend of the epidemic. On average, wastewater data anticipated reported cases by about 9 days, providing an early warning of an increase in cases in the health context of the study.

We have also proposed a simple theoretical framework that integrated wastewater data into a compartmental epidemic model. This framework allowed to estimate the actual prevalence of infection, which was found to be about 53%, compared to the 19% detected in the same period in Catalonia. This discrepancy suggests that there was a large and time-variable under-reporting in the detection of infections, especially at the onset of the epidemic. We argued that this under-reporting was fundamentally related to people's perception of the epidemic state and the information available to them, generating a vicious circle.

We have estimated the maximum quantity of genome copies shed in a gram of feces by an individual during the course of the infection, which results to be between 1.4×10^8 gc/g and 4.4×10^8 gc/g, which showed a good agreement with the literature.

Other works tried to incorporate wastewater data into analytical models; for instance, Cavany et al. (2022), analyzing wastewater data from a university campus, inferred that wastewater data are highly overdispersed, highlighting therefore the limits of wastewater surveillance as a leading indicator, in contrast with what emerged in this study. However, the context of their study differs relevantly from that of the present work: they analyzed wastewater data from a small university campus during fall 2020 consisting of mainly young people who were isolated from the campus as long as they were identified as infected, resulting in a completely different epidemiological framework. Moreover, they assumed a shedding profile that peaks 6 days after infection and a mean incubation period of only 2 days, in contrast with the assumptions of this work and with those of many other studies. This probable overestimation of the time for the peak of the shedding
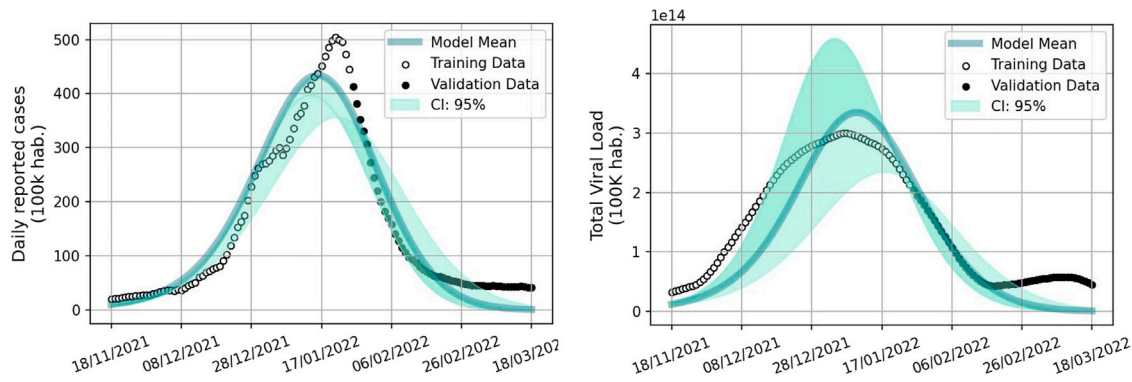
Fig. 5. Model validation and spatiotemporal propagation of COVID-19 across Catalonia visualized through daily reported cases and total viral load in wastewater plants. The figure shows the comparison between model predictions and data about daily reported cases (left side) and total quantity of genome copies of SARS-CoV-2 in sewage. Solid lines show model predictions for the daily reported cases (left) and the daily number of genome copies in sewage (right) for 100,000 inhabitants, whereas dots correspond to real data. The model has been trained for the first 100 days data (white dots) and validate in the remaining ones (black dots). The shadowed areas represent the 95% prediction interval. The R^2 statistics is 0.94 for cases and 0.8 for wastewater data.
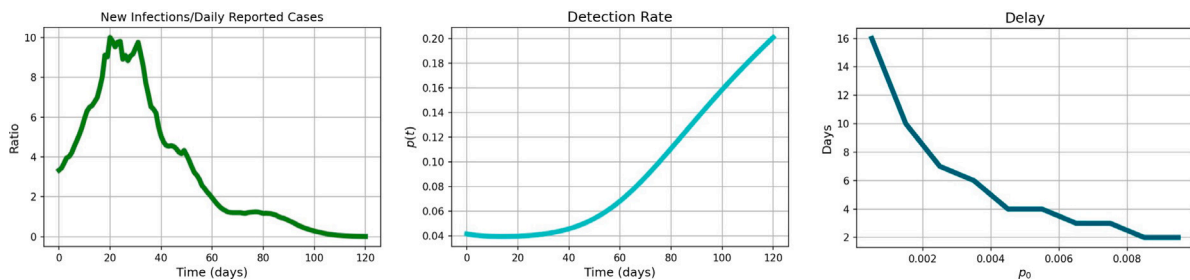
Fig. 6. Under-reporting (left panel), detection rate (central panel) and delay versus p_0 (right panel). Left panel: temporal evolution of the ratio between new infections simulated by the model at each time step and daily reported cases data. Central panel: temporal evolution of the detection rate $p(t)$ according to the model estimates. Right panel: days of delay between generated quantity of genome copies concentrations and detected infections varying the value of p_0. The former is deduced again looking to the maximum Pearson correlation between the two simulated data-sets of genome copies and detected infections.
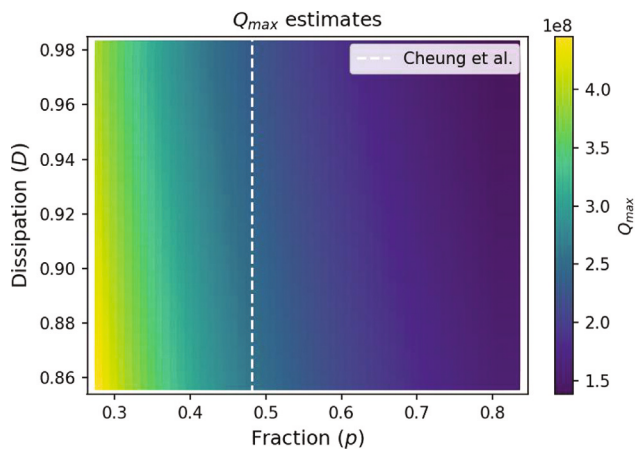

Fig. 7. Values of Q_{max} varying the dissipation factor D and the fraction p and inferred by the estimated value for $\bar{k}$. Colormap showing estimations for the maximum quantity of genome copies shed in feces by an individual Q_{max} inferred from Eq. (2), varying the dissipation factor (D) and the fraction of people shedding virus in feces (p), according to the indications in the literature. We used the mean value of $\bar{k}$ in the posterior distribution yielded by the model calibration procedure. The white dashed line indicates the value for p suggested by Cheung et al. (2020) in their review. .

process and the underestimation of the incubation period do not allow to appreciate the effectiveness of wastewater data as leading indicator. Another assumption of this work was to consider a constant detection rate in time for reported cases; in this paper we showed how the temporal variability of the detection rate can largely affect the delay

between wastewater data and reported cases curves, and therefore the potential for the former to be used as leading indicator.

In their work, Ando et al. (2023) developed a new sensitive method for longitudinal tracking of C_{RNA} and a mathematical model allowing for the prediction of COVID-19 cases from wastewater data. They employed the shedding profile proposed by Miura et al. (2021) whose peak is in the early phase of the infection as in our framework. However they calibrated their model using reported cases as trustworthy estimator of the real prevalence of infections, without considering any biases in the clinical surveillance. We also used the combined information from reported cases and wastewater data but also allowing for a differentiation between real number of infections versus the detected ones.

Here we want to remark few aspects that can limit our analysis. The main limitation is represented by the data itself: as pointed out by Huisman et al. (2022), with less than three samples per week the measurements of genome copies in sewage can change according to the day of the data taking. We are looking forward to improve our weekly data-set increasing the number of samples per week.

In general, we are aware that a more complex model is needed to model SARS-CoV-2 epidemic involving other aspects like mobility, protection measures, restrictions, age stratification etc. and, in particular, to express such intricate concept like the people perception and awareness about the epidemic. Indeed, other aspects of human behavior can be taken in consideration, as imitation processes or adoption of different strategies, given that human behavior and epidemic spreading undergo to a complex interaction that goes in both directions. We are also aware that mechanistic models that try to in-globe wastewater data cannot be extremely accurate, due to the intrinsic volatility and the multitude of factors that enter in the entire process of the virus shedding in the sewage system. For instance, Thomas et al. (2017) highlighted how considering dynamical populations, for which the

number of persons served by a specific sewage plant can change in time, is way more accurate than fixed ones. Morvan et al. (2022) showed how machine learning models result naturally more accurate in capturing the wastewater phenomenology. We also considered only fecal excretion in the model, ignoring other forms of viral shedding like through saliva. However, we analyzed wastewater concentrations for municipalities whose population ranges from 10.000 to over 1 million inhabitants and, according to Crank et al. (2022), in large communities stool shedding dominates by far over the other possible forms. Furthermore, the SIR model and its basic variations do not consider re-infections or the onset of other variants, therefore the epidemic wave naturally dies out after a certain time. This can limit our analysis because the period from October 2021 to March 2022 corresponds to the transition from VOC Delta to VOC Omicron BA.1, then BA.2.

5. Conclusions

Considering the challenges mentioned, the modeling framework developed in this work should not be viewed as a real-world application model for highly accurate predictions, if such a thing is even possible. Instead, our model's primary goal is to capture the general epidemic trend suggested by wastewater data and reported cases, with a focus on evaluating the effectiveness of clinical diagnostic tests in the context of the study.

The benefit of wastewater-based data is that it avoids the systematic error present in reported cases when used as an estimator for actual prevalence, as it inevitably misses numerous infected individuals who go untested for various reasons. Although wastewater data are inherently noisy, they do not exhibit this selection bias, but rather a more stochastic imprecision.

As a result, we think that our theoretical framework offers valuable insights and a general approach for integrating wastewater data into compartmental models. We aimed to investigate what we consider the two main questions of WBE: (i) our work indicates that wastewater-based data can serve as a leading indicator signal, not only due to the shedding profile peaking before symptom onset, but also because the lead time is ultimately influenced by the manner in which clinical tests are conducted, with low and variable detection rates affecting this delay; (ii) even though wastewater data cannot provide highly accurate prevalence estimations due to natural randomness and modeling complexity, incorporating this new source of information as a complementary tool within simple models can emphasize the significant underestimation of infections by clinical diagnostic tests in contexts similar to that of this study.

CRediT authorship contribution statement

Mattia Mattei: Designed the study, Analyzed the data, Performed the analysis, Analyzed the results, Wrote the paper, Reviewed. **Rosa M. Pintó:** Provided the data, Reviewed. **Susana Guix:** Provided the data, Reviewed. **Albert Bosch:** Provided the data, Reviewed. **Alex Arenas:** Designed the study, Analyzed the results, Wrote the paper, Reviewed.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Ahmed, W., Angel, N., Edson, J., et al., 2020. First confirmed detection of sars-cov-2 in untreated wastewater in australia: A proof of concept for the wastewater surveillance of covid-19 in the community. Sci. Total Environ. 728 (138764).

Ando, Hiroki, Murakami, Michio, Ahmed, Warish, Iwamoto, Ryo, Okabe, Satoshi, Kitajima, Masaaki, 2023. Wastewater-based prediction of covid-19 cases using a highly sensitive sars-cov-2 rna detection method combined with mathematical modeling. Environ. Int. 173, 107743.

Arenas, Alex, Cota, Wesley, Gómez-Gardeñes, Jesús, Gómez, Sergio., Granell, Clara, Matamalas, Joan T., Soriano-Paños, David., Steinegger, Benjamin, 2020. Modeling the spatiotemporal epidemic spreading of covid-19 and the impact of mobility and social distancing interventions. Phys. Rev. X 10, 041055.

Arts, P., Banman, B., Anglin, K., Kelly, D., Boehm, A., Wolfe, M., Wigginton, K., 2022. Estimating sewershed prevalence of sars cov 2, pmmov, and crass phage fecal shedding using models informed by measurements from covid 19 patient samples and wastewater solids. In: Virus Outbreaks, Epidemiology and Source Tracking, ISFEV 2022.

Bivins, Aaron, Greaves, Justin, Fischer, Robert, Yinda, Kwe Claude, Ahmed, Warish, Kitajima, Masaaki, Munster, Vincent J., Bibby, Kyle, 2020. Persistence of sars-cov-2 in water and wastewater. Environ. Sci. Technol. Lett. 7 (12), 937–942.

Carcereny, A., Garcia-Pedemonte, D., Martínez-Velázquez, A., Quer, J., Garcia-Cehic, D., Gregori, J., Antón, A., Andrés, C., Pumarola, T., Chacón-Villanueva, C., Borrego, C.M., Bosch, A., Guix, S., Pintó, R.M., 2022b. Dynamics of SARS-CoV-2 Alpha (B.1.1.7) variant spread: The wastewater surveillance approach. Environ Res. 208, 112720. http://dx.doi.org/10.1016/j.envres.2022.112720.

Carcereny, A., Martínez-Velázquez, A., Bosch, A., Allende, A., Truchado, P., Cascales, J., Romalde, J.L., Lois, M., Polo, D., Sánchez, G., Pérez-Cataluña, A., Díaz-Reolid, A., Antón, A., Gregori, J., Garcia-Cehic, D., Quer, J., Palau, M., Ruano, C.G., Pintó, R.M., Guix, S., 2021. Monitoring Emergence of the SARS-CoV-2 B.1.1.7 Variant through the Spanish National SARS-CoV-2 Wastewater Surveillance System (VATar COVID-19). Environ Sci Technol. 55 (17), 11756–11766. http://dx.doi.org/10.1021/acs.est.1c03589.

Carcereny, Albert, et al., 2022a. Dynamics of sars-cov-2 alpha (b.1.1.7) variant spread: The wastewater surveillance approach. Environ. Res. 208, 112720.

Cavany, Sean, Bivins, Aaron, Wu, Zhenyu, North, Devin, Bibby, Kyle, Alex Perkins, T., 2022. Inferring sars-cov-2 rna shedding into wastewater relative to the time of infection. Epidemiol. Infect. 150, e21.

Chavarria-Miro, G., Anfruns-Estrada, E., Martinez-Velazquez, A., Vazquez-Portero, M., Guix, S., Paraira, M., Galofre, B., Sanchez, G., Pinto, R.M., Bosch, A., 2021. 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. Appl. Environ. Microbiol. 87 (7).

Cheung, K.S., Hung, I.F.N., Chan, P.P.Y., et al., 2020. Gastrointestinal manifestations of sars-cov-2 infection and virus load in fecal samples from a hong kong cohort: Systematic review and meta-analysis. Gastroenterology 159 (1), 81–95.

Crank, K., Chen, W., Bivins, A., Lowry, S., Bibby, K., 2022. Contribution of sars-cov-2 rna shedding routes to rna loads in wastewater. Sci. Total Environ. 806, 150376.

D'Aoust, P.M., Graber, T.E., Mercier, E., et al., 2021. Catching a resurgence: Increase in sars-cov-2 viral rna identified in wastewater 48 h before covid-19 clinical tests and 96 h before hospitalizations. Sci. Total Environ. 770, 145319.

Fernandez-Cassi, Xavier, Scheidegger, Andreas, Banziger, Carola, Cariti, Federica, Corzon, Alex Tunas, Ganesanandamoorthy, Pravin, Lemaitre, Joseph C., Ort, Christoph, Julian, Timothy R., Kohn, Tamar, 2021. Wastewater monitoring outperforms case numbers as a tool to track covid-19 incidence dynamics when test positivity rates are high. Water Res. 200, 117252.

Hoffmann, Till, Alsing, Justin, 2021. Faecal shedding models for sars-cov-2 rna amongst hospitalised patients and implications for wastewater-based epidemiology. medRxiv.

Huisman, Jana S., Scire, J., et al., 2022. Wastewater-based estimation of the effective reproductive number of sars-cov-2. Environ. Health Perspect. 150 (15).

Kermack, W.O., McKendrick, A.G., 1927. A contribution to the mathematical theory of epidemics. Proc. R. Soc. Lond. Ser. A 115 (772), 700–721.

Krivoakova, N., Soltysova, A., Tamas, M., et al., 2021. Mathematical modeling based on rt-qpcr analysis of sars-cov-2 in wastewater as a tool for epidemiology. Sci. Rep. 11, 19456.

Larsen, D.A., Wigginton, K.R., 2020. Tracking covid-19 with wastewater. Nat. Biotechnol. 38, 1151–1153.

Marjoram, P., Molitor, J., Plagnol, V., Tavare, S., 2003. Markov chain monte carlo without likelihoods. Proc. Natl. Acad. Sci. 100 (26), 15324–15328.

McCall, Camille, Fang, Zheng N., Li, Dongfeng, Czubai, Andrew J., Juan, Andrew, LaTurner, Zachary W., Ensor, Katherine, Hopkins, Loren, Bedient, Philip B., Stadler, Lauren B., 2022. Modeling sars-cov-2 rna degradation in small and large sewersheds. Environ. Sci.: Water Res. Technol. 8, 290–300.

McMahan, C.S., Self, S., Rennert, L., et al., 2021. Covid-19 wastewater epidemiology: a model to estimate infected populations. Lancet Planet Health 5 (12), e874–e881.

Miura, F., Kitajima, M., Omori, R., 2021. Duration of sars-cov-2 viral shedding in faeces as a parameter for wastewater-based epidemiology: Re-analysis of patient data using a shedding dynamics model. Sci. Total Environ. 769, 144549.

Moreno, J.L., Zuniga, S., Enjuanes, L., Sola, I., 2008. Identification of a coronavirus transcript enhancer. J. Virol. 82, 3882–3893.

Morvan, M., Jacomo, A.L., Souge, C., et al., 2022. An analysis of 45 large-scale wastewater sites in england to estimate sars-cov-2 community prevalence. Nature Commun. 13, 4313.

Olesen, S.W., Imakaev, M., Duvallet, C., 2021. Making waves: Defining the lead time of wastewater-based epidemiology for covid-19. Water Res. 202, 117433.

Peccia, J., Zulli, A., Brackney, D.E., et al., 2020. Measurement of sars-cov-2 rna in wastewater tracks community infection dynamics. Nat. Biotechnol. 38, 1164–1167.

Phan, Tin, Brozak, Samantha, Pell, Bruce, Gitter, Anna, Xiao, Amy, Mena, Kristina D., Kuang, Yang, Wu, Fuqing, 2023. A simple seir-v model to estimate covid-19 prevalence and predict sars-cov-2 transmission using wastewater-based surveillance data. Sci. Total Environ. 857, 159326.

Poletti, Piero, Ajelli, Marco, Merler, Stefano, 2011. The effect of risk perception on the 2009 h1n1 pandemic influenza dynamics. PLOS ONE 6 (2), 1–7.

Randazzo, W., Truchado, P., Cuevas-Ferrando, E., Simon, P., Allende, A., Sanchez, G., 2020. Sars-cov-2 rna in wastewater anticipated covid-19 occurrence in a low prevalence area. Water Res. 181 (115942).

Steinegger, Benjamin, Arenas, Alex, Gómez-Gardeñes, Jesús, Granell, Clara, 2020. Pulsating campaigns of human prophylaxis driven by risk perception palliate oscillations of direct contact transmitted diseases. Phys. Rev. Res. 2, 023181.

Teunis, P.F., Sukhrie, F.H., Vennema, H., Bogerman, J., Beersma, M.F., Koopmans, M.P., 2015. Shedding of norovirus in symptomatic and asymptomatic infections. Epidemiol. Infect. 143 (8), 1710–1717.

Thomas, K.V., Amador, A., Baz-Lomba, J.A., Reid, M., 2017. Use of mobile device data to better estimate dynamic population size for wastewater-based epidemiology. Environ. Sci. Technol. 51 (19), 11363–11370.

Toni, Tina, Welch, David, Strelkowa, Natalja, Ipsen, Andreas, Stumpf, Michael P.H., 2009. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. J. R. Soc. Interface 6 (31), 187–202.

Vemulapalli, R., Gulani, J., Santrich, C., 2009. A real-time taqman rt-pcr assay with an internal amplification control for rapid detection of transmissible gastroenteritis virus in swine fecal samples. J. Virol. Methods 162, 231–235.

Wallis, C., Melnick, J.L., 1967. Concentration of viruses on aluminum and calcium salts. Am. J. Epidemiol. 85, 459–468.

Wang, W., Xu, Y., Gao, R., Lu, R., Han, K., Wu, G., Tan, W., 2020. Detection of sars-cov-2 in different types of clinical specimens. JAMA 323 (18), 1843–1844.

Weidhaas, J., Aanderud, Z.T., Roper, D.K., VanDerslice, J., Gaddis, E.B., Ostermiller, J., Hoffman, K., Jamal, R., Heck, P., Zhang, Y., Torgersen, K., Laan, J.V., LaCross, N., 2021. Correlation of sars-cov-2 rna in wastewater with covid-19 disease burden in sewersheds. Sci. Total Environ. 775 (145790).

Wolfel, R., Corman, V.M., Guggemos, W., et al., 2020. Virological assessment of hospitalized patients with covid-2019. Nature 581, 465–469.

Wu, Yongjian, Guo, Cheng, Tang, Lantian, Hong, Zhongsi, Zhou, Jianhui, Dong, Xin, Yin, Huan, Xiao, Qiang, Tang, Yanping, Qu, Xiujuan, Kuang, Liangjian, Fang, Xiaomin, Mishra, Nischay, Lu, Jiahai, Shan, Hong, Jiang, Guanmin, Huang, Xi, 2020a. Prolonged presence of sars-cov-2 viral rna in faecal samples. Lancet Gastroenterol. Hepatol. 5 (5), 434–435.

Wu, F., Xiao, A., Zhang, J., et al., 2020b. Sars-cov-2 titers in wastewater foreshadow dynamics and clinical presentation of new covid-19 cases. medRxiv.

Xu, Y., Li, X., Zhu, B., et al., 2020. Characteristics of pediatric sars-cov-2 infection and potential evidence for persistent fecal viral shedding. Nat. Med. 26 (502–505).

Zhang, Ning, Gong, Yuhuan, Meng, Fanping, Bi, Yuhai, Yang, Penghui, Wang, Fusheng, 2020. Virus shedding patterns in nasopharyngeal and fecal specimens of covid-19 patients. medRxiv.