Environmental Technology & Innovation 36 (2024) 103720

Contents lists available at ScienceDirect

Environmental Technology & Innovation

journal homepage: www.elsevier.com/locate/eti

Multi-factor normalisation of viral counts from wastewater improves the detection accuracy of viral disease in the community

Cameron Pellett [a], Kata Farkas [a,*], Rachel C. Williams [a], Matthew J. Wade [b], Andrew J. Weightman [c], Eleanor Jameson [a], Gareth Cross [d], Davey L. Jones [a]

[a] School of Environmental and Natural Sciences, Bangor University, Bangor, Gwynedd LL57 2UW, UK
[b] Data, Analytics & Surveillance Group, UK Health Security Agency, London E14 4PU, UK
[c] Cardiff School of Biosciences, Cardiff University, Cardiff CF10 3AX, UK
[d] Science Evidence Advice Division, Health and Social Services Group, Welsh Government, Cathays Park, Cardiff, CF10 3NQ, UK

ARTICLE INFO

Keywords:
Viral pandemics
Disease prevalence
Machine learning
Predictive modelling
Wastewater-based epidemiology

ABSTRACT

The detection of viruses (e.g. SARS-CoV-2, norovirus) in wastewater represents an effective way to monitor the prevalence of these pathogens circulating within the community. However, accurate quantification of viral concentrations in wastewater, proportional to human input, is constrained by a range of uncertainties, including (i) dilution within the sewer network, (ii) degradation of viral RNA during wastewater transit, (iii) catchment population and facility use, (iv) efficiency of viral concentration and extraction from wastewater, and (v) inhibition of amplification during the RT-qPCR step. Here, we address these uncertainties by investigating several potential normalisation factors including the concentration of ammonium and orthophosphate. A faecal indicator virus (crAssphage), and the recovery of the process-control viruses (murine norovirus and bacteriophage Phi6), used for quality control during the RT-qPCR step, were also considered. We found that multi-factor normalisation of SARS-CoV-2 RT-qPCR data was optimal using a combination of crAssphage, process-control virus recovery, and concentration efficiency to improve prediction accuracy relative to clinical test data. Using multi-normalised SARS-CoV-2 RT-qPCR data, we found a lasso regression model with random forest modelled residuals lowers the prediction error of positives by 46 %, compared to a single linear regression using raw data. This multi-normalised approach enables more accurate wastewater-based predictions of clinical cases up to five days in advance of clinical data, identifying trends in disease prevalence before clinical testing, and demonstrates the potential to improve viral pathogen detection for a range of currently monitored and emerging diseases.

1. Introduction

The COVID-19 pandemic necessitated the development of new methods for early detection and surveillance of SARS-CoV-2 as infections emerged in communities. This need drove rapid innovation in wastewater-based epidemiology (WBE), leading to the growth and refinement of techniques to detect pathogenic viruses at urban wastewater treatment plants (WWTPs; Shah et al., 2022; Parkins et al., 2023). WBE can be used to identify areas with growing disease prevalence allowing the timely implementation of mitigation

* Corresponding author.
E-mail address: k.farkas@bangor.ac.uk (K. Farkas).

https://doi.org/10.1016/j.eti.2024.103720
Received 12 February 2024; Received in revised form 31 May 2024; Accepted 13 June 2024
Available online 29 June 2024

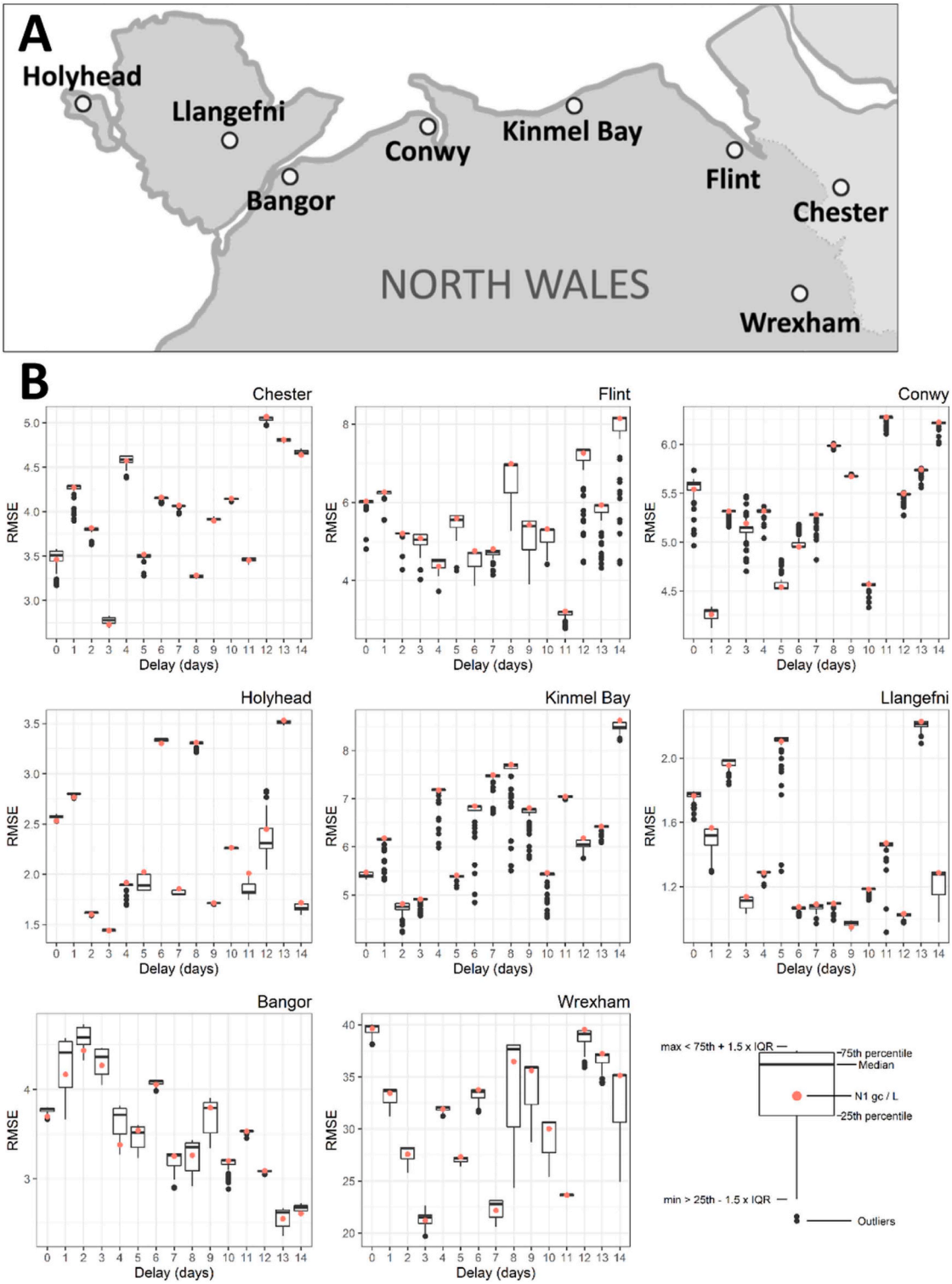

Fig. 1. A. Map of North Wales and a section of England showing the location of the Wastewater treatment plants (WWTPs) sites. B. Lagged correlation between normalised and unnormalised wastewater SARS-CoV-2 measures and positive case rates. They depict the root-mean-square-error (RMSE) of single linear models predicting COVID-19 case rates with 79 normalised and unnormalised wastewater SARS-CoV-2 N1 gene fragment quantities, predicting zero to fourteen days ahead of clinical test data. The best time lag between the clinical signal and the wastewater signal was selected by the lowest median RMSE (Chester: 3 days; Flint: 11 days; Conwy: 1 day; Holyhead: 3 days; Kinmel Bay: 2 days; Llangefni: 9 days; Bangor: 13 days; Wrexham: 3 days).

measures aimed at controlling further spread of the disease (Bittihn et al., 2021; Henriques et al., 2023). This has proved useful as individual testing often fails to detect all infections due to asymptomatic cases, pre-symptomatic transmission, limited access to clinical testing facilities or poor uptake due to testing hesitancy (McGowan et al., 2020; Bourrier and Deml, 2022). The approach has now been applied to a wide range of enteric and respiratory pathogens including SARS-CoV-2, norovirus, influenza, *Campylobacter* and mpox (formerly known as monkeypox) (Wannigama et al., 2023; Zhang et al., 2023; Shrestha et al., 2024). WBE measures of viral RNA/DNA in wastewater typically enables: (i) detection prior to symptom onset, (ii) unbiased sampling of the whole community, including asymptomatic cases, (iii) rapid, inexpensive deployment (Buscarini et al., 2020; Gibas et al., 2021; Kumar et al., 2021; Shah et al., 2022; Shrestha et al., 2021; Wade et al., 2022), and (iv) advanced warning of rising prevalence due to high asymptomatic rates, allowing early localised response (Jones et al., 2020; Gibas et al., 2021; Zhu et al., 2021a).

In the case of COVID-19, efforts to develop effective concentration methods for isolation of viral RNA in wastewater has allowed for the quantification of the SARS-CoV-2 N1 gene fragment with reverse transcription quantitative polymerase chain reaction (RT-qPCR) (Farkas et al., 2021; Ahmed et al., 2021). Targeting the N1 gene fragment has been particularly successful in detecting the virus in sewage, up to three weeks in advance of clinical cases, but also in predicting absolute case rates at a community level (Gibas et al., 2021; Ahmed et al., 2021; D'Aoust et al., 2021; Karthikeyan et al., 2021). However, the latter requires accurate quantification proportional to the levels of virus shed by infected individuals. The levels of virus detected, however, are directly influenced by a range of factors including: (i) individual shedding rate, (ii) dilution in wastewater by rainfall and other inputs, (ii) decay of the RNA/DNA during transit in the sewer network or post-sample collection, (iii) catchment population and facility use, and (iv) efficiency of concentration, extraction and RT-qPCR of genetic material in the laboratory (Ahmed et al., 2020; Wade et al., 2022). However, once the viral concentrations are determined, data should be adjusted to population levels in the WWTP catchment area.

Chemical analytes, including ammonium (NH_4^+), orthophosphate (PO_4^{3-}), cotinine and 5-hydroxyindoleacetic acid, have been linked to census populations and employed to normalise measured levels of illicit drugs in wastewater (Been et al., 2014; Van Nujis et al., 2011; Chen et al., 2014), with ammonium linked to daily and hourly fluctuations due to commuter, weekend and other travel dynamics (Been et al., 2014). More recently, viral population markers have been adopted, including pepper mild mottle virus (PMMoV), F+ coliphage and crAssphage, all of which are prevalent in the human gut and indicative of human faeces in water (Kitajima et al., 2018; Farkas et al., 2019; Rainey et al., 2023). These viruses, therefore, offer the possibility to normalise SARS-CoV-2 levels in wastewater, with evidence that this reduces exogenous variability due to flow and results in improved correlations with positive clinical tests (Wu et al., 2020; D'Aoust et al., 2021; Wilder et al., 2021).

Additionally, process control viruses, added to the samples at different stages of sample processing, can be used to estimate viral recoveries. For example, mengovirus (a Picornavirus infecting mainly rodents) has been used to normalise norovirus concentration for inhibition during RT-qPCR of environmental samples (Haramoto et al., 2018; Zakhour et al., 2010). However, most studies only attempt to incorporate one or two variables for normalisation (Bertels et al., 2023; Sweetapple et al., 2023), and none have attempted to normalise for inhibition and viral loss during concentration, extraction and RT-qPCR simultaneously, leaving a significant amount of variability unaccounted for and a considerable gap for improvements in normalisation.

In this work, our aims were to (a) assess the correlation between wastewater measures and clinical case rates, (b) further analyse crAssphage, ammonium and orthophosphate as normalisation variables for wastewater dilution, viral decay, catchment population and facility/building-scale use, (c) introduce and assess the crAssphage concentration factor and process-control virus recovery as normalisation variables for efficiency of concentration, extraction and RT-qPCR, (d) evaluate the influence of viral shedding profiles on prediction accuracy, and (e) expand the statistical modelling tools used beyond linear and parametric regression. To do this, we carried out a longitudinal study, measuring wastewater data before, during and after the second peak of the COVID-19 epidemic from September 2020 to March 2021, for eight cities and towns in the UK.

2. Materials and methods

2.1. COVID-19 clinical case data

Clinical COVID-19 case data that had been geographically aggregated to lower layer super output areas (LSOAs) were obtained from Public Health Wales. LSOAs are defined by the national population census data as regions comprising between 400 and 1200 households equating to a resident population of between 1000 and 3000 persons (ONS, 2021). Cases were then assigned to WWTPs, using Arc-GIS (Esri Inc., Redlands, CA), by the ratio of the LSOA's population serviced by the WWTP's catchment area.

2.2. Sampling methodology

Untreated influent wastewater ($n = 572$) was collected 2–5 times per week using grab samples between 25th September 2020 and 22nd March 2021 at eight municipal WWTPs located at Chester, Wrexham, Flint, Kinmel Bay, Conwy, Bangor, Llangefni and Holyhead in North Wales and Northwest England, UK (Fig. 1A). Wastewater samples (500 ml) were collected plant-side of the crude influent grate in sterile Nalgene® bottles in the morning between 08.00 h and 10.00 h, at the time of peak flow following the protocol of Hillary et al. (2021). This type of sampling has been shown to give very similar results to composite sampling using refrigerated autosamplers at the same locations (Farkas et al., 2023). Samples were stored and transported to the laboratory at 4°C and processed within 24 h of collection.

2.3. Chemical analysis

The pH and electrical conductivity (EC) of the wastewater samples were measured using a Hanna 209 pH meter (Hanna Instruments Ltd., UK) and a Jenway 4520 conductivity meter (Jenway Ltd., UK). Prior to ammonium and orthophosphate determination, samples were centrifuged (18,000 *g*, 15 min) to remove suspended solids. After recovery of the supernatant, the concentrations of ammonium and orthophosphate were determined colorimetrically using the salicylic acid procedure of Mulvaney (1996) and ammonium molybdate method of Murphy and Riley (1962), respectively. In brief, samples were mixed with ammonium molybdate, sulfuric acid and ascorbic acid and absorbance were measured at 820 nm. Orthophosphate concentration was determined against a dilution series of PO_4-P standards. For determining ammonium concentrations, samples were mixed with EDTA tetrasodium salt dihydrate, sodium nitroprusside and NaK_2PO_4. Once green colour appeared, absorbance was measured at 667 nm. Ammonium concentration was determined against a dilution series of NH_4^+-N standards. All measurements were conducted using a Spectrostar Nano microplate reader (BMG Labtech, Germany).

2.4. Sample concentration and viral RNA/DNA extraction

A 200 ml aliquot of each sample was centrifuged to eliminate solid matter. A 0.2–0.5 ml aliquot of the clarified supernatant was retained for direct nucleic acid extraction using the NucliSense MiniMag extraction system (BioMerieux, France) for crAssphage quantification (hereafter referred to as the unconcentrated supernatant). Subsequently, viral genetic material in 150 ml of the supernatant was concentrated using the polyethylene glycol (PEG) 8000 methods, Fig. S1, described in Farkas et al. (2021). The resulting PEG pellet was either resuspended in 0.5 ml phosphate buffered saline (PBS; pH 7.4) and the viral nucleic acids were extracted manually using the NucliSense MiniMag extraction system (BioMerieux, France) or the PEG pellet was resuspended in 0.8 ml NucliSense Lysis buffer and extracted using the NucliSense chemistry on the Kingfisher 96 Flex extraction system (Thermo Scientific) as previously described (Kevill et al., 2022). The final volume of the eluents was 0.1 ml.

The samples taken between 13th January 2021 and 22nd March 2021 were spiked with $\sim 10^5$–10^6 genome copies (gc) of an enveloped dsRNA bacteriophage, the *Pseudomonas* phage Phi6 (DSM 21518; Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, Germany; Fedorenko et al., 2020) after the initial centrifugation step. The Phi6 phage was cultured in a *Pseudomonas* sp. host (DSM 21482) in Tryptone Soy broth (TSB) containing 5 µl 0.5 M $CaCl_2$ at 25 °C (Kevill et al., 2022). With each set of samples, a process control (200 ml ion-exchanged water spiked with Phi6) was also processed to assess concentration efficiency and cross-contamination.

The samples taken between 25th September 2020 and 22nd January 2021 were spiked with $\sim 10^5$–10^6 genome copies of an ssRNA virus, murine norovirus (MNV), after PEG precipitation for extraction efficiency control. The MNV was cultured and prepared in BV2 cells, kindly provided by Prof Ian Goodfellow, University of Cambridge, UK, as previously described (Kevill et al., 2022). In brief, MNV was cultured in BV2 tissue in high-glucose Dulbecco's minimum essential medium (DMEM; Sigma Aldrich, USA) with 10 % foetal bovine serum (FBS; Sigma Aldrich, USA), 1 % MEM Non-Essential Amino Acids (Sigma Aldrich, USA), 1 % L-glutamine (Sigma Aldrich, USA) and 1 % Penicillin/Streptomycin (Sigma Aldrich, USA) at 37 °C with 5 % CO_2 for 48 h. After cytopathic effect was confirmed, viral particles were extracted by three freeze-thaw cycles followed by centrifugation (1000 *g*, 15 min). The supernatant was then diluted 10 × in PBS (pH 7.4) and stored at − 80 °C. With each set of extraction, negative control (PBS) and positive control (PBS spiked with MNV) were used to assess extraction efficiency and cross-contamination.

2.5. RNA/DNA quantification of viral targets

All real-time PCR assays were run on the QuantStudio Flex 6 system (Applied Biosystems, USA) and analysed using the QuantStudio real-time PCR software v1.3 and v1.7 (Applied Biosystems, USA). In all assays, samples were run in duplicate and molecular-grade water was used as a non-template control to assess cross-contamination. The primers and probes used in the assays are listed in Table S1 and the standard curve details are listed in Table S2. Samples were subject to (RT-)qPCR within 24 h after RNA/DNA extraction. Altogether 192 (RT-)qPCRs were run.

We quantified the faecal indicator, crAssphage, in the samples using a previously described assay with Quantifast Probe reaction mix (Qiagen, Germany) with the reaction mix and conditions as described previously (Farkas et al., 2019; Kevill et al., 2022). We used a dilution series of plasmid DNA incorporating the target sequence as standards. For SARS-CoV-2 N1, MNV and Phi6 quantification, we used one-step qRT-PCR either with the RNA Ultrasense 1-step qRT-PCR reaction mix (Applied Biosystems, USA) (Farkas et al., 2021) or with the TaqMan Virus Fast 1-step qRT-PCR mix (Applied Biosystems, USA) (Kevill et al., 2021). For standards, a dilution series (10^0–10^5 gc/µl) of synthetic RNA were used in duplicates (Kevill et al., 2022). Where applicable, the SARS-CoV-2 N1 and Phi6 assays were duplexed by mixing the primers and probes of the two assays in one reaction mix. When duplexed assayed were applied, a singular standard dilution series with both target RNA sequences was prepared. Duplexing did not affect assay performance (Farkas et al., 2022; Kevill et al., 2022).

The concentration data was expressed as genome copies (gc)/µl nucleic acid extract. Sample concentrations (gc/l wastewater) were calculated as in Eq. (1).

$$vc_w = \frac{vc_e \bullet ev}{sv} \bullet 1000 \quad (1)$$

Where vc_w is the virus concentration in a wastewater sample (gc/l); vc_e is the virus concentration in RNA eluent (gc/µl); ev is the RNA eluent volume (µl); and sv is the processed sample supernatant volume (ml).

2.6. Statistical analysis

Statistical analysis was carried out in R v4.1.2, utilising the "dplyr", "purr", "tidyr", "slider" and "zoo" packages for data manipulation; the "ggplot2", "gridExtra", "ggpubr" and "patchwork" packages for visualisations; and the "glmnet" and "randomForest" packages for lasso regression and random forest models, respectively (R Core Team, 2021; Wickham, 2016, 2020; Wickham et al., 2021; Henry and Wickham, 2020; Grolemund and Wickham, 2011; Vaughan, 2020; Zeileis and Grothendieck, 2005; Auguie, 2017; Kassambara, 2020; Pedersen, 2020; Friedman et al., 2010; Liaw and Wiener, 2002).

2.7. Viral sample processing controls

To account for variation in efficiency of concentration and extraction, a concentration factor was calculated from the measurement of crAssphage with and without sample concentration (Eq. (2)). Essentially, if concentration has reduced performance, potentially due to chemical interference, the crAssphage measured in the concentrated sample (C) will be proportionally lower, whereas the crAssphage measured in the unconcentrated supernatant sample (S) will remain stable, thus reducing the concentration factor (CF).

$$CF = \frac{C}{S} \tag{2}$$

To account for variation in concentration, extraction and qPCR efficiency, the process control virus (Phi6 or MNV) recovery was calculated through comparison of control virus quantified in wastewater samples with control virus quantified in sample controls (Eq. (3)). For example, if the qPCR was inhibited by an unmeasured chemical in the wastewater sample, the quantity of control virus measured in the wastewater sample would reduce compared to the quantity of control virus measured in the control, resulting in a decreased control virus recovery.

$$cv_{rec} = \frac{cv_s}{cv_c} \tag{3}$$

Where cv_{rec} is the Phi6 or MNV control virus recovery calculated as the proportion of control virus quantified in a wastewater sample (cv_s) to control virus quantified in a control (cv_c).

Due to the change in internal control from MNV to Phi6, missing data in MNV or Phi6 recovery (MNV_{rec}; $Phi6_{rec}$) were estimated using a linear model predicted with the control used (Eqs. (4a) and (4b)). The model was fit using overlapping data during the transition period.

$$Phi6_{rec} = \beta_0 + \beta_1 \bullet MNV_{rec} + \varepsilon \tag{4a}$$

$$MNV_{rec} = \beta_0 + \beta_1 \bullet Phi6_{rec} + \varepsilon \tag{4b}$$

Where β_0 and β_1 are estimated with QR factorization, optimising the residual-sum-of-squares (RSS; Eq. (5)) of the model. While ε is the error remaining in the model.

$$SS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \; x_{ij} \right)^2 \tag{5}$$

Where y_i is observation i of the variable being predicted and x_{ij} is observation i of the predictor variable j; β_0 is the estimated model intercept; and β_j is the estimated parameter j for variable j.

2.8. Normalisation of SARS-CoV-2 concentrations in wastewater

Calculated variables were assigned to categories for influencing factors they should account for and used to normalise the SARS-CoV-2 estimations (Fig. S2). Normalisation was carried out by multiplying viral concentrations by normalisation coefficients for each variable (Table S3). Normalisation coefficients are calculated simply as the reciprocal of each of the processing controls described above. Calculation of the crAssphage concentrate normalisation coefficient (C_{nc}) is demonstrated in Eq. (6a). A combined mean normalisation coefficient ($comb_{nc}$) is shown in Eq. (6b) that takes the average of the crAssphage (C), ammonium (NH_4) and orthophosphate (P) normalisation coefficients but represents the only combination of normalisation variables of this kind.

$$C_{nc} = \frac{1}{C} \tag{6a}$$

$$comb_{nc} = \left(\frac{1}{P} + \frac{1}{NH4} + \frac{1}{C} \right) \Big/ 3 \tag{6b}$$

Normalisation coefficients were then used to normalise SARS-CoV-2 N1 concentration (gc/l), as demonstrated with Phi6 recovery in Eq. (7), normalising N1 gene fragment abundance for inhibition of concentration, extraction and qPCR.

$$N1_{norm} = N1 \bullet \frac{1}{Phi6_{rec}} = \frac{N1}{Phi6_{rec}} \tag{7}$$

Further, normalisation coefficients were combined to account for many factors at once, as demonstrated with pre-normalised crAssphage (C) by MNV in Eq. (9), normalising for dilution, viral decay, facility use and inhibition during extraction. Additionally, square root transformations were performed on the normalisation variables (Eq. (9)) to mitigate the impact of large outliers created due to the grab sampling approach taken (Wade et al., 2022). Logarithmic transformations were considered but not used, to avoid interference with the assessment of the shedding profile (discussed in Section 2.10).

$$N1_{norm} = \frac{N1}{C/\sqrt{MNV}} \tag{8}$$

In total, 79 variables were created through a variety of singular and combined normalisations (full list shown in Table S3).

2.9. Delayed correlation between SARS-CoV-2 in wastewater and clinical case data

Analysis was independently performed on each of the eight sites, following the same methodology. To assess the time lag between wastewater measurements and clinically derived case rates in the population, linear models were fitted for each of the 79 normalised variables, predicting case rates between zero to fourteen days ahead. The 1185 models for each site were then evaluated using the root mean square error (RMSE; Eq. (10)), and the best fitting time lag was selected as the delay with the lowest median RMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n}} \tag{9}$$

Where y_i is the case rates; $\widehat{y_i}$ is the predicted case rates; and n is the total number of predictions.

2.10. Normalisation variable assessment

After selecting and applying the best fitting lag (Section 2.8), the impact of each normalisation variable and their combinations were assessed. The models were assigned to groups based on the variables and combinations used in normalisation, then visualised and compared with raw, unnormalised SARS-CoV-2 N1 gene fragment abundance (gc/l) based on their RMSE. Additionally, the RMSE for the models was normalised to the case rate standard deviation of each site (NRMSE; Eq. (11)) to allow for a combined comparison. Welch's two sample t-tests were selected, following an assessment of distribution, to compare crAssphage pre-normalisation and square root transformation.

$$NRMSE = RMSE \Big/ \frac{\sqrt{\sum_{i=1}^{n}(y_i - \mu)^2}}{n} \tag{10}$$

Where: μ is the mean case rate.

2.11. Human shedding profile of SARS-CoV-2

To assess the potential impact of viral shedding profiles on the prediction of case rates using wastewater measurements, a decay rate (D) was formulated as shown in Eq. (12), assuming a logistic decay for the shedding profile shown by Hoffmann and Alsing (2023), and calculated with half-lives ranging from 0 to 100 h.

$$D = 0.5^{t - t_{-1}/half-life} \tag{11}$$

The decay rate (D) was applied to SARS-CoV-2 N1 estimations, as shown in Eq. (13), to calculate the expected concentration of virus from newly infected individuals ($N1_{new}$) after accounting for continued shedding from previously infected and positively tested individuals. Furthermore, the decay rate was applied to the best performing normalised N1 variable, with performance defined by prediction accuracy (RMSE) of case rates using a linear model.

$$N1_{new} = N1_t - N1_{t_{-1}} \bullet D \tag{12}$$

2.12. Improved predictive modelling

A lasso regression model with random forest modelled residuals was selected as a final prediction of case rates, with the lasso regression optimised by minimising the ℓ_1 penalised residual sum of squares (ℓ_1RSS; Eq. (13)), and the random forest optimised by minimising the residual sum of squares (RSS; Eq. (5)). The model was selected due to the ability of linear models to extrapolate beyond the bounds of their training data, lasso models having inherent robustness against multicollinearity, and the combination of parametric

and non-parametric approaches allowing linear and non-linear interactions to be predicted (Zhang et al., 2019).

$$\ell_1 RSS = RSS + \lambda \sum_{j=1}^{p} |\beta_j| \tag{13}$$

Where n is the total number of observations; p the total number of variables; and λ is an adjustable tuning parameter.

The lasso regression tuning parameter (λ) was selected through k-fold cross-validation (CV) ($k = 10$; Jung, 2018), where the dataset is split into k subsets and k models can be fit with one subset removed for each model, such that every subset of data is omitted from model fitting in one model. The models can then each be tested on their omitted subset to produce a more reliable estimate of model performance when applied to new data. As such, an optimal λ was selected for reducing the CV error, with a larger λ resulting in greater shrinkage of model coefficients, and vice versa.

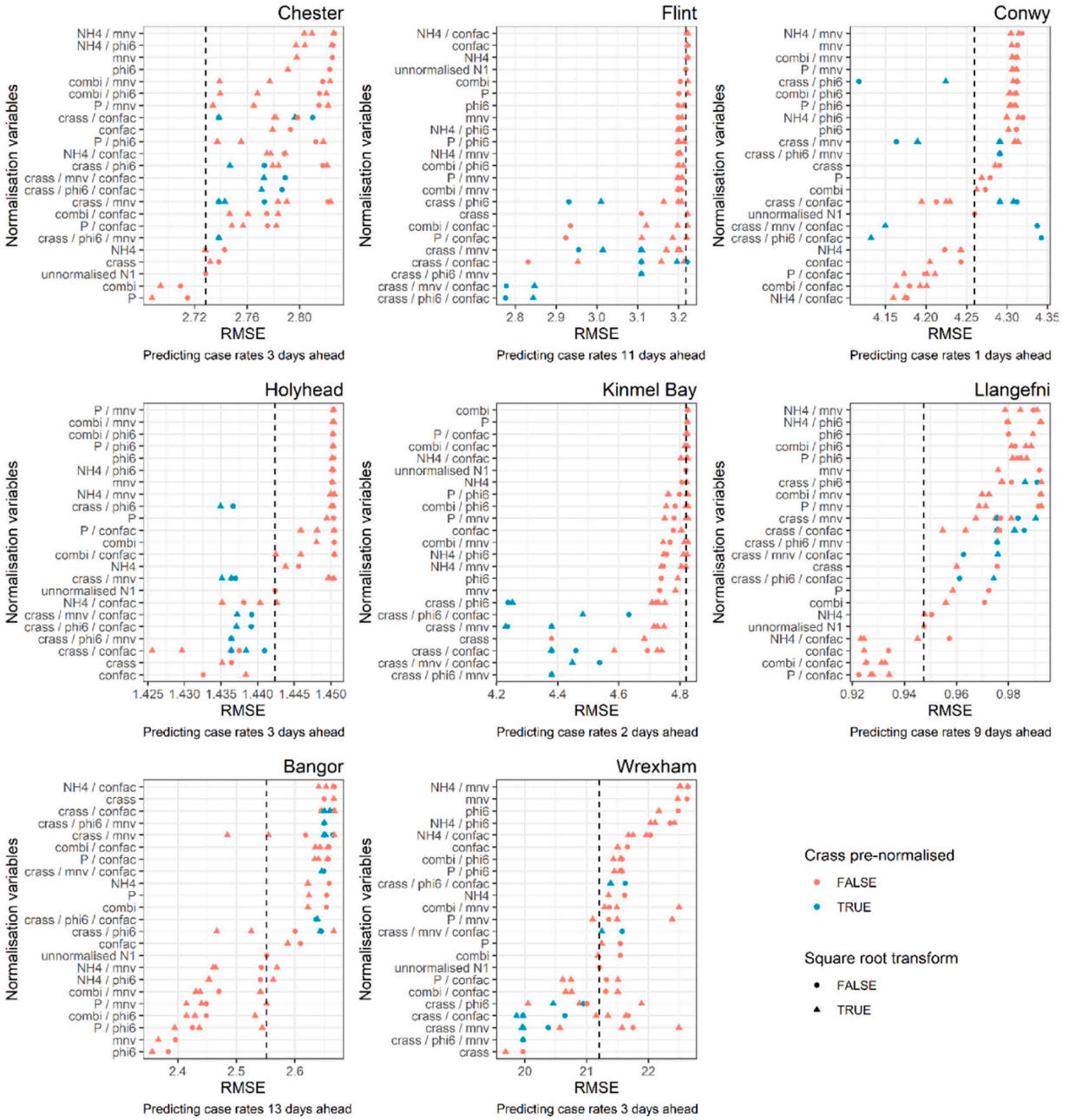

Fig. 2. Comparison of normalisation variables. Depicts the RMSE of single linear models predicting case rates using a variety of normalisation variables, at the best delayed correlation. The dashed line is centred on the unnormalised N1 gene fragment abundance, identifying normalisations which improved the model fit with a lower RMSE to its left.

The random forest was then fit to the lasso regression residuals, calculated as the observed minus predicted cases rates. A random forest is a collection of t regression trees, fitted using v randomly selected variables at each node. Where t was selected through k-fold CV of random forest models with 115–450 trees at 15 tree intervals, and v was selected as the square root of the number of initial variables. The initial variables for the random forest with normalised variables were selected through backward propagation, removing 20 % at each iteration with the lowest importance, then the best subset of variables was selected through k-fold CV. Importance was calculated as the total RSS (Eq. (5)) reduced through branches selected with each variable. Two lasso regressions with random forest modelled residuals were fitted: the first without normalised variables and the second with all normalised variables prior to initial variable selection, as detailed above.

The model fit was assessed using leave-one-out CV, and calculation of the CV RMSE. Leave-one-out CV functions exactly as k-fold, except k is equal to the number of observations, thus retaining more data in the fitting process with the constraint of being more computationally demanding. Models for each site were fit with the wastewater parameters predicting case rates zero to fourteen days into the future. Then, after selecting the optimum delay, the final models were fit, and the cross-validated predictions visualised against the true clinical case rates.

3. Results

3.1. Wastewater measures correlate with future clinical cases

Three temporal variables (wastewater transport time, pre-symptomatic shedding and community reactions to testing) have conflicting impacts on the delay between wastewater sample collection and the onset of symptoms and positive clinical tests. To account for this, single linear models were fit to clinical positive tests, zero to fourteen days ahead, with each of the 79 variables created from normalisation of the SARS-CoV-2 N1 gene fragment (see Section 2.7 for details; variable list in Supplementary Table S3; Fig. S3-5 for heatmaps; and Table S4 for summary data). The 1185 models for each site were then grouped by delay and the lowest median root-mean-square-error (RMSE) was used to select the optimum. This optimum lag time had a mode of three days, although this varied considerably between sites (from 1 to 13 days; Fig. 1). This was determined based on the lowest median RMSE overall. Selection of the best lag time would not have varied if based on unnormalised SARS-CoV-2 N1 measures, or the single best variable, rather than the median. However, there were no clear downward and upward trends around an optimum, rather, obvious individual optima surrounded by noise, or sometimes no clear optimum.

3.2. Multi-normalisation improves correlation with case rates

Variation in wastewater dilution, viral decay, population numbers and sample process efficiency could have significant impacts on the validity and accuracy of predicted disease prevalence using WBE. To assess this, we grouped the models by the variable(s) used to normalise N1 gene fragment concentration, identifying the best combination by the minimum RMSE or RMSE normalised by the sites'

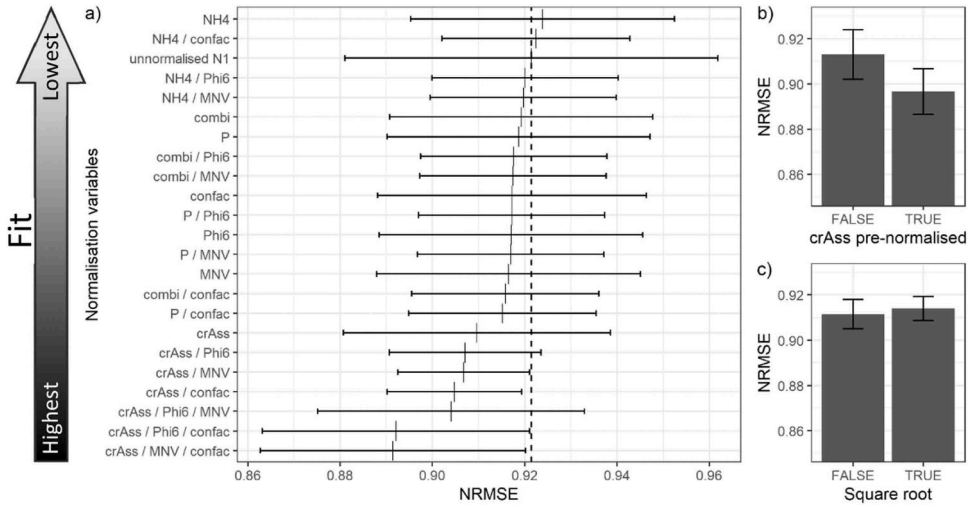

Fig. 3. Combined site and delay comparison of normalisation variables. a) Depicts the normalised-root-mean-square-error (NRMSE; RMSE normalised by the site standard deviation) of each combination of normalisation variables for all sites and all delays. The centre lines indicate the mean NRMSE and the tails indicate the 95 % confidence intervals with a t distribution. The dashed line is centred on the mean unnormalised SARS-CoV-2 N1 gene fragment abundance, identifying variables that have a greater likelihood of improving the model fit with a lower NRMSE to its left, as indicated by the Fit arrow to the left of the y-axis. b) NRMSE of crAssphage (crAss) normalisation variables with and without pre-normalisation where the bar indicates the mean, and tails indicate the 95 % confidence intervals with a t distribution. c) NRMSE of all normalisation variables with and without square root transformation.

standard deviation (NRMSE) when multiple sites were compared together. First, we assessed normalisation variables using data from the best selected delay by individual sites (Fig. 2, see Table S3 for full list of normalisation variables). Then, we assessed normalisation variables using data with all sites and delays combined (Fig. 3). All normalisation variables were useful for some sites, however, after selecting for the best delay, their effectiveness varied greatly: some sites (e.g. Chester and Llangefni) saw improvements with very few variables, while others (e.g. Flint and Kinmel Bay) saw improvements with almost all. Notably, crAssphage, and particularly crAssphage combined with Phi6/MNV recovery and concentration factor, had the greatest likelihood of improving prediction accuracy through normalisation (Fig. 3a). In contrast, orthophosphate concentration combined with MNV and the combination of crAssphage, orthophosphate and ammonium (combi) still gave some improvements to the model. Thus, crAssphage or combined crAssphage (combi) normalisations improved prediction accuracy over unnormalised N1 gene fragment abundance at all eight sites (Fig. 2). Additionally, on average, pre-normalisation of crAssphage improved prediction accuracy, but the means could not be assumed different (Welch's two sample t-test: $t = 1.84$, $df = 3062.4$, p-value $= 0.065$; Fig. 3b). In comparison, square root transformation reduced prediction accuracy but, similarly, the means could not be assumed to be different (Welch's two sample t-tests: $t = -0.53$, $df = 9464$, p-value $= 0.593$; Fig. 3c). Despite this, the single best predictor had square root transformations in five of the eight sites (Fig. 2).

3.3. *Shedding from infected individuals follows logistic decay, but does not influence wastewater monitoring performance*

Prolonged shedding of virus RNA from previously infected individuals combined with current positively tested individuals has been suggested to influence the use of WBE for SARS-CoV-2 by overestimation (Hoffmann and Alsing, 2023; Puhach et al., 2023). To evaluate the effect of post-symptomatic faecal shedding, we applied a logistic decay to normalised and unnormalised N1 gene fragment values, with half-lives between 0 and 100 hours, predicting the expected quantity at the following timestep if there were no new infections. This continued shedding is subtracted from the value measured in the wastewater (see Section 2.10 for details). A logistic

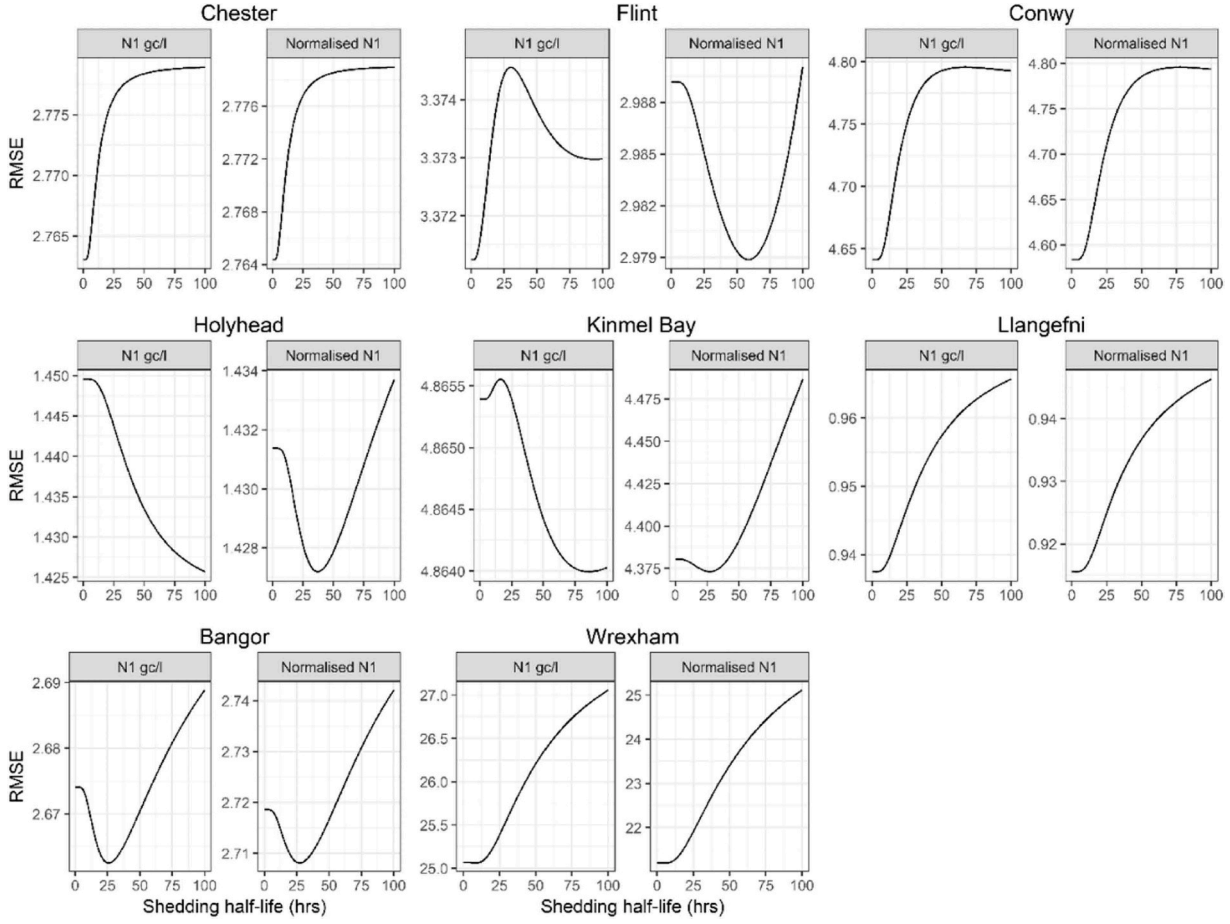

Fig. 4. RMSE of linear models predicting case rates with normalised and unnormalised wastewater variables with an assumed logistic decay shedding profile with half-lives between 0 and 100 h. The normalised variables had a mean optimum half-life of 20.0 h, whilst the unnormalised variable had a mean optimum of 28.2 h. Comparing the RMSE without decay and with the optimum decay had a mean improvement of 0.004 with normalised variables and 0.005 with unnormalised SARS-CoV-2 N1 gene fragment abundance values.

decay in faecal shedding profile was supported by an improved RMSE in seven of the eight sites, with an optimum half-life of 20.0 h for normalised variables, and 28.2 h for unnormalised N1 gene fragment abundance, with the best delay selected (Fig. 4). The normalised variables improved by a mean RMSE of 0.004 cases after decay was incorporated, and the unnormalised N1 gene fragment abundance improved by 0.005 cases (Fig. 4). Shedding profile by individual sites varied considerably, with three showing minimal improvement ($<$ 0.001 RMSE) with a decay function applied (e.g. Conwy, Llangefni and Wrexham), and one saw reduced performance with an increased RMSE (Chester) (Fig. 4). However, of the sites which were improved considerably using the decay function, the optimum half-life generally fell between 15 and 60 h, for normalised variables (Fig. 4).

3.4. Machine learning and normalisation allow accurate prediction of future clinical cases

To assess the prediction improvements investigated through normalisation and advanced machine learning approaches, we fitted four sets of models to each site, predicting clinically derived case rates at the best selected delay based on the leave-one-out cross-validated RMSE. The first and second set of models were fit using single linear regression with unnormalised SARS-CoV-2 N1 gene fragment abundance and the best normalised N1 gene variable, respectively (see Table S5 for best variable). For the third and fourth, we fitted lasso regression models with random forest modelled residuals, re-optimising the best time delay, and utilising only unnormalised variables for the third and a subset of all normalised and unnormalised variables for the fourth. Normalisation improved the cross-validated prediction accuracy in all sites, and the use of lasso regression with random forest modelled residuals with normalised variables improved the prediction accuracy over all other models, reducing the mean RMSE from unnormalised N1 by 46 % (Table 1). Furthermore, pairwise comparisons using Wilcoxon signed rank exact tests suggest models one, two and four were all significantly different from one another, as well as models three and four (p-value $<$ 0.05 with the Holm-Bonferroni (H-B) adjustment method; Holm, 1979). However, model three was not significantly different from models one and two (p-value $>$ 0.05 with the H-B adjustment method).

The fit of the fourth and final model accounted for most of the main trends, particularly in Flint, Conwy and Wrexham - the three sites with the highest COVID-19 case rates (Fig. 5; mean daily case rate $>$ 8). However, some variation was still missed, largely in sites with lower case rates, including Chester, Bangor and Llangefni, suggesting the need for additional variables. These results suggest positive clinical case rates can be predicted up to fourteen days ahead with reasonable accuracy (mean RMSE: 4.16 cases; Table 1). Although, with the inaccuracy of some of the smaller sites, a prediction of up to five days is likely more reasonable (Fig. 5). Estimates from the lasso regression indicated an optimal model for Holyhead, Kinmel Bay, Llangefni and Bangor with only an intercept, suggesting no wastewater variable improved the linear fit (Fig. 6a). Wrexham, Flint and Conwy, on the other hand, were optimised with multiple variables including N1 normalised with pre-normalised crAssphage by concentration factor, and non-N1 variables including ammonium level and concentration factor. Additionally, Chester utilised a single variable (N1 normalised by orthophosphate) albeit with a very small coefficient, reduced by the ℓ_1 penalised regression (Eq. (13)). The random forest fit to the lasso regression residuals found greater non-parametric importance in wastewater variables (Fig. 6b). SARS-CoV-2 N1 gene fragment abundance normalised by crAssphage was used for all sites and the most important variable for six sites, which was combined with or pre-normalised by Phi6 or concentration factor for four sites. Phi6 or MNV, was utilised in all sites, with each being used in six sites. Orthophosphate and concentration factor were also important, each being used for five sites, whilst NH_4^+-N was less important, being utilised for three sites. These results further support the use of multiple normalisation variables to improve predictions of COVID-19 prevalence in communities.

4. Discussion

The protocol described here combined multi-normalisation and machine learning to generate accurate wastewater-based

Table 1
Model comparison based on leave-one-out cross-validated RMSE. Comparing single linear regression to a lasso regression with random forest residuals; model 1 consisted of unnormalised SARS-CoV-2 N1 gene fragment values, model 2 used the best normalised N1 gene variable, model 3 used only unnormalised variables, and model 4 was a subset of all normalised and unnormalised variables.

WWTP site	Leave-one-out cross-validated RMSE			
	1 Linear regression N1 gc/l	2 Linear regression best normalised N1	3 Lasso with random forest (not normalised)	4 Lasso with random forest (normalised)
Chester	3.09	2.81	2.94	2.65
Flint	4.03	3.48	3.7	2.99
Conwy	10.4	4.79	4.71	3.22
Holyhead	1.51	1.50	1.84	1.46
Kinmel Bay	8.14	4.35	5.13	4.38
Llangefni	0.97	0.97	1.15	0.87
Bangor	3.07	2.96	2.82	2.57
Wrexham	30.4	24.8	18.3	15.1
Mean ± SEM	7.70 ± 3.43	5.71 ± 2.76	5.07 ± 1.94	4.16 ± 1.61

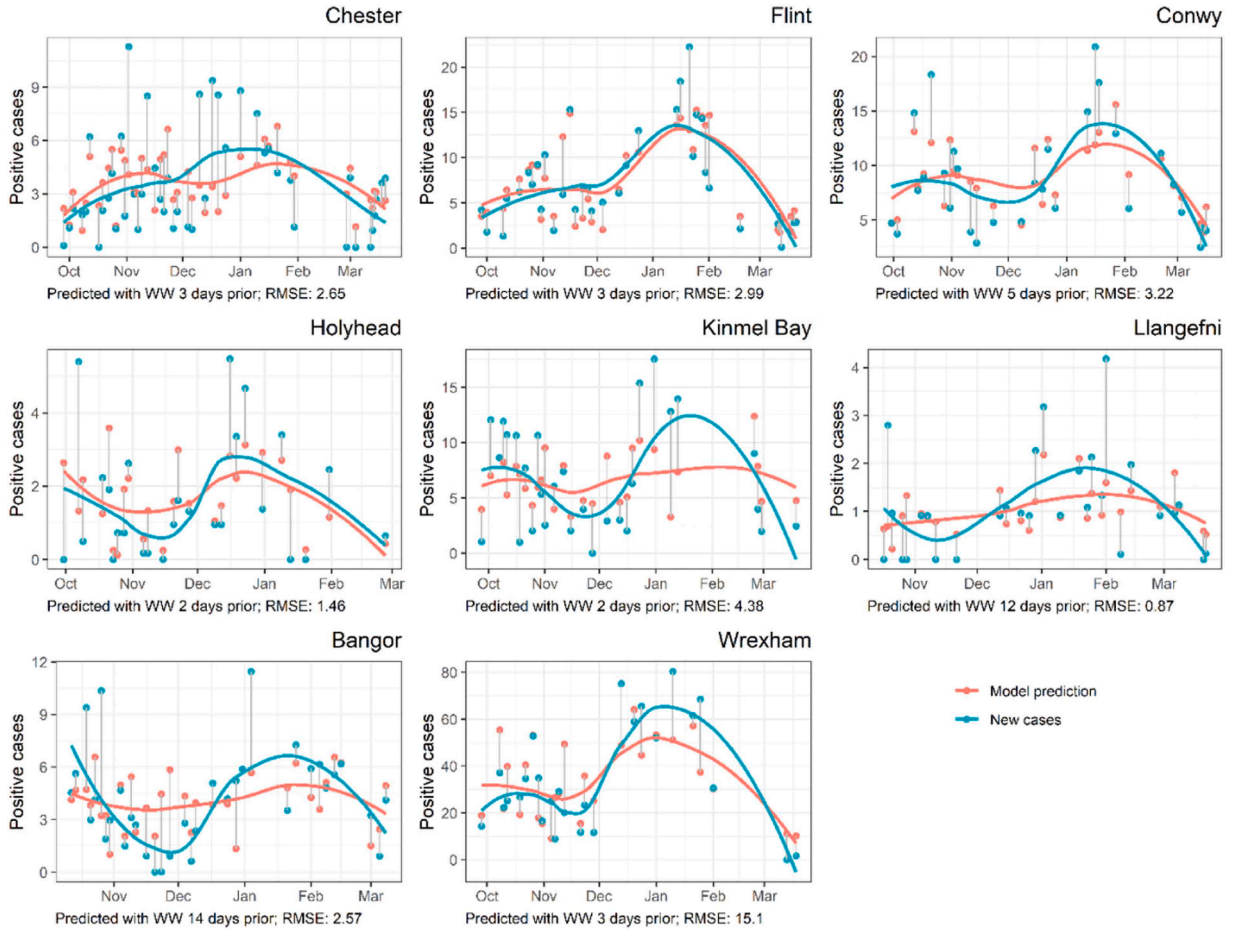

Fig. 5. Cross-validated results from a lasso regression with random forest modelled residuals predicting case rates with a variety of normalised and unnormalised wastewater variables (Fig. 6). Red points indicate the cross-validated model predictions and blue dots indicate true individual testing cases rates from positive clinical PCR tests. A LOESS trend line was fitted to both the model prediction and true case rates to identify general trends and assess model fit. Models were fit for Chester ($n = 47$), Flint ($n = 37$), Conwy ($n = 31$), Holyhead ($n = 26$), Kinmel Bay ($n = 29$), Llangefni ($n = 30$), Bangor ($n = 37$), and Wrexham ($n = 26$) individually.

predictions of clinical viral infections within 30 h of sampling, comparable to clinical testing turnaround times 24–48 h (Larremore et al., 2021). Importantly, however, the wastewater COVID-19 signal occurred days in advance of clinical presentation, which enables mitigation options to be imposed earlier and preparations to be made in advance of patient arrival in healthcare facilities. The potential of WBE to identify disease outbreaks earlier and more cost-effectively than clinical testing has previously been demonstrated (Gibas et al., 2021; Buscarini et al., 2020). However, wastewater detection is impacted by many more confounding factors than clinical sampling, and accurate viral quantification and modelling is challenging (Jiang et al., 2023). Our results show that wastewater detection of SARS-CoV-2, normalised with crAssphage and a process control virus reduced the error of positive tests by 46 % and is reliable for cities and large towns to predict future case rates.

While our approach involves testing multiple normalizations, the final models are optimized for each site individually. This site-specific optimization is needed to account for the unique characteristics and variability of each sewershed (e.g. Wade et al., 2022; Jiang et al., 2023). Our analysis provides clear evidence of a lag time between clinical data and wastewater data of ca. 3 days, consistent with delays in clinical case data reported in other work (Wilder et al., 2021; D'Aoust et al., 2021; Hillary et al., 2021). This acknowledges that the delay in clinical detection was site-specific. We ascribe site specificity to differences in transport times, community behaviour, and interactions with the vaccine in different locations, meaning that delay factor should be optimised for each site. As the virus shedding profile may vary during the course of infection, we investigated the effect of introducing virus half-lives into our normalisation (Natarajan et al., 2022). Previous work monitoring SARS-CoV-2 RNA in hospitalised patients' stool samples suggested a logistic decay after peak shedding (Hoffmann and Alsing, 2023). A faecal shedding half-life of 34 h was found in these hospitalised patients, which more closely matches unnormalised SARS-CoV-2 N1 gene fragment abundance, however it is unlikely that the response of hospitalised patients would match that of the general population, where a more rapid decay in shedding profile could be expected (Hoffmann and Alsing, 2023). Our results suggest that continued shedding over days and weeks will not have a considerable impact on SARS-CoV-2 in wastewater and, therefore, the decay was not applied in further assessment and modelling.

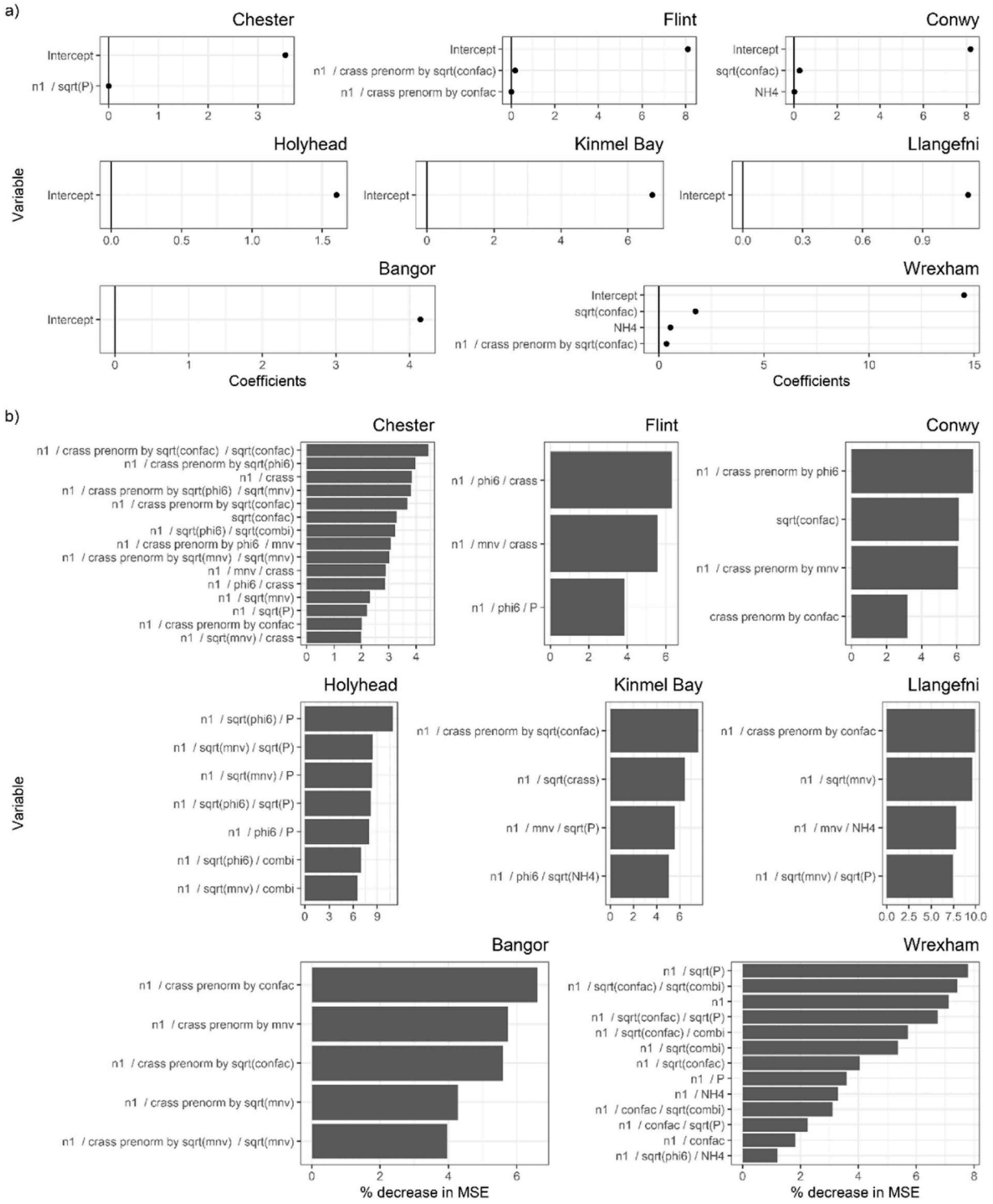

Fig. 6. Variable selection and importance for the lasso regression, with random forest modelled residuals fit to each site at the best selected delay for the final model. a) Lasso regression non-zero variable coefficients. b) Random forest variable importance indicated by the percentage decrease in mean-squared-error (MSE) attributed to each variable. Crass: crassphage, n1: SARS-CoV-2, NH_4: ammonium, P: orthophosphate.

In this study, we utilised normalization factors, such as dilution, viral decay, population dynamics, and sample processing efficiency, which have been shown to significantly impact the accuracy of WBE (Feng et al., 2021; Hsu et al., 2022; Langeveld et al., 2023). These variables were consistently important across our sites, suggesting their value in accounting for population dynamics and sample

processing efficiency. Interestingly, the faecal indicator virus crAssphage proved to be the best normalisation variable. This was despite its previously stated difficulties in measuring viruses accurately in wastewater (Haramoto et al., 2018). Our results are supported by previous studies assessing crAssphage normalisation, which found it resulted in better correlations to clinical test data (Wilder et al., 2021; Greenwald et al., 2021). However, due to variations in crAssphage abundance in the population, crAssphage should only be used at sites serving more than 5600 people (Langeveld et al., 2023). Potentially, variability introduced by PCR inhibition during the measurement of crAssphage DNA is proportional to that for SARS-CoV-2 RNA. Thus, after normalisation, crAssphage may simultaneously adjust for population faecal input relative to dilution, as with ammonium and orthophosphate, and PCR inhibition. Further improvements by normalising with the process control virus, Phi6, could then be explained by the viral genome structure due to differential interactions between chemical inhibitors and the non-enveloped DNA virus crAssphage, versus the enveloped RNA viruses Phi6 and SARS-CoV-2 (Gendron et al., 2010; Stachler et al., 2017). The combined use of DNA/RNA faecal indicators have shown to improve normalisation previously (Mitranescu et al., 2022) as well as in this study. Interestingly, previous studies have suggested that normalisation with process control virus (bovine coronavirus) and with faecal indicators, such as pepper mild mottle virus (PMMoV) and HF183 gene, reduced SARS-CoV-2 wastewater signal correlation with case data (Feng et al., 2021; Hsu et al., 2022; Maal-Bared et al., 2023). This implies that the efficiency of faecal indicators and virus recovery methods may vary, highlighting the importance of careful selection.

Ammonium and orthophosphate normalisation produced site dependent mixed results, and were not selected in the final models for every site, potentially due to competition during model fitting with crAssphage. These wastewater indicators have previously been used successfully for normalisation (Been et al., 2014; Van Nuijs et al., 2011), reflecting the complexity of wastewater and inherent variability between sites. Other chemicals, such as pharmaceuticals or lifestyle-associated markers, and water flow may also be useful for normalisation (Hsu et al., 2022; Kasprzyk-Hordern et al., 2023; Langeveld et al., 2023; Maal-Bared et al., 2023; Mitranescu et al., 2022). As such, all normalisation variables should be assessed in further studies as the use of multiple normalisation variables in the final models improved accuracy significantly.

Despite the improvements in accuracy achieved in this study, it is potentially limited by the sampling methodology and frequency. Data used in this study was generated from wastewater collected by grab sampling, which captures a sample at a single time point and may not be representative of the general shedding patterns over the day (Gerrity et al., 2021; Wu et al., 2020; Wade et al., 2022). The use of 24-hour composite sampling can provide a more representative data point and, thus, may reduce model by lowering measurement variance (Gerrity et al., 2021; Wu et al., 2020). However, this is not conclusive and a previous study indicates that is may provide negligible improvement (Farkas et al., 2023). Regardless, the effectiveness of normalisation using composite SARS-CoV-2 measurements remains unresolved. Secondly, due to limited access to WWTPs during the pandemic, sampling frequency was inconsistent during the study period, particularly at Kinmel Bay and Wrexham. This inconsistency may impact selection of the optimal lag time between the wastewater and clinical signal. In addition to these limitations, the methodology is still impacted by inherent variability of equipment and detection limits, which still require considerable development, as addressed in other work (Zhu et al., 2021b; El Soufi et al., 2024).

To prevent overestimation of model accuracy due to overfitting, we employed leave-one-out cross-validation to assess model performance. Cross-validation assesses how well models generalise to unseen data and provides a realistic estimate of a model's predictive performance (Zhang et al., 2019). Furthermore, we utilized lasso regression, which performed feature selection and regularization, to mitigate overfitting and enhance model interpretability (Tibshirani, 1996).

More generally, the methodology presented has broad applications beyond its use for SARS-CoV-2, primarily with other emerging viral pathogens. Furthermore, normalisation with process controls and factor calculations could be used to assess any quantified genomic data and other analytes in WWTPs, particularly those requiring concentration or purification steps. Broader still, multi-normalisation prior to machine learning, when using algorithms robust to multi-collinearity (e.g. random forest and lasso regression), could result in considerable improvements in accuracy for many alternate applications. These include the prediction of water sources from downstream samples based on physiochemical and microbial data, and identifying potential functions for new drugs based on physicochemical and structural attributes (Wang et al., 2021; Vamathevan et al., 2019). Both Wang et al. (2021) and Vamathevan et al. (2019) utilised random forest algorithms, but neither combined this with multi-normalisation. Other recent studies on COVID-19 WBE also utilised random forest models, however, without normalising for sample process efficiency (Dejus et al., 2023; Li et al., 2023). Our work improved these applications with the use of a single focal independent variable showing that exhaustive normalisation with population and sample process efficiency factors could be carried out, where every variable is manipulated by one another in every unique combination, prior to variable subset selection. The broader applications of this work pave the way for improvements in many other fields.

5. Conclusions

- Results suggest a common two to five-day delay between the wastewater signal and an increase in clinical positive tests.
- Multi-comparison of normalisation factors for the SARS-CoV-2 N1 gene target found that normalisation by crAssphage, process control virus recovery and concentration factor were the variables with the highest likelihood of improving prediction accuracy.
- Our evidence suggests that continued shedding of SARS-CoV-2 from previously infected individuals does not significantly influence WBE for clinical case rate estimation.
- The use of multi-normalisation and a lasso regression model with random forest modelled residuals reduced the prediction error of positive tests by 46 %, allowing major trends to be predicted, up to five days in advance.

- The lasso regression and random forest identified crAssphage normalisations as the most frequently and important, followed by process control virus recovery, orthophosphate concentration and concentration factor. NH_4^+-N normalisations were less important, but may still improve normalisation at some sites.
- The approach described here integrates multi-factor normalization, process controls, and advanced machine learning techniques to improve the accuracy and reliability of WBE.

CRediT authorship contribution statement

Kata Farkas: Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization. **Cameron Pellett:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Eleanor Jameson:** Writing – review & editing, Writing – original draft, Methodology. **Andrew J. Weightman:** Writing – review & editing, Writing – original draft, Funding acquisition. **Matthew J Wade:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Rachel C Williams:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation. **Davey L Jones:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization. **Gareth Cross:** Writing – review & editing, Methodology, Funding acquisition.

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.

Acknowledgements

This work was funded by UK Research and Innovation (UKRI) under the COVID-19 Rapid Response Programme (Projects NE/V004883/1 and NE/V010441/1), the Centre for Environmental Biotechnology Project funded though the European Regional Development Fund (ERDF) by Welsh Government and the Welsh Government COVID-19 surveillance programme. We particularly thank Steve Cobley at Welsh Government and Tony Harrington and staff at Dŵr Cymru Welsh Water. We thank Prof Ian Goodfellow (University of Cambridge, UK) for providing MNV and BV2 stocks.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.eti.2024.103720.

References

Ahmed, W., Bivins, A., Bertsch, P.M., Bibby, K., Choi, P.M., Farkas, K., Gyawali, P., Hamilton, K.A., Haramoto, E., Kitajima, M., Simpson, S.L., Tandukar, S., Thomas, K., Mueller, J.F., 2020. Surveillance of SARS-CoV-2 RNA in wastewater: methods optimization and quality control are crucial for generating reliable public health information. Curr. Opin. Environ. Sci. Health 17, 82–93. https://doi.org/10.1016/j.coesh.2020.09.003.

Ahmed, W., Tscharke, B., Bertsch, P.M., Bibby, K., Bivins, A., Choi, P., Clarke, L., Dwyer, J., Edson, J., Nguyen, T.M.H., O'Brien, J.W., 2021. SARS-CoV-2 RNA monitoring in wastewater as a potential early warning system for COVID-19 transmission in the community: a temporal case study. Sci. Total Environ. 761, 144216 https://doi.org/10.1016/j.scitotenv.2020.144216.

Auguie, B., 2017. gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. ⟨https://CRAN.R-project.org/package=gridExtra⟩.

Been, F., Rossi, L., Ort, C., Rudaz, S., Delémont, O., Esseiva, P., 2014. Population normalization with ammonium in wastewater-based epidemiology: application to illicit drug monitoring. Environ. Sci. Technol. 48, 8162–8169. https://doi.org/10.1021/es5008388.

Bertels, X., Hanoteaux, S., Janssens, R., Maloux, H., Verhaegen, B., Delputte, P., Boogaerts, T., van Nuijs, A.L., Brogna, D., Linard, C., Marescaux, J., 2023. Time series modelling for wastewater-based epidemiology of COVID-19: a nationwide study in 40 wastewater treatment plants of Belgium, February 2021–June 2022. Sci. Total Environ. 899, 165603.

Bittihn, P., Hupe, L., Isensee, J., Golestanian, R., 2021. Local measures enable COVID-19 containment with fewer restrictions due to cooperative effects. EClinicalMedicine 32, 100718. https://doi.org/10.1016/j.eclinm.2020.100718.

Bourrier, M.S., Deml, M.J., 2022. The legacy of the pandemic preparedness regime: an integrative review. Int. J. Public Health 67, 1604961. https://doi.org/10.3389/ijph.2022.1604961.

Buscarini, E., Manfredi, G., Brambilla, G., Menozzi, F., Londoni, C., Alicante, S., Iiritano, E., Romeo, S., Pedaci, M., Benelli, G., Canetta, C., La Piana, G., Merli, G., Scartabellati, A., Viganò, G., Sfogliarini, R., Melilli, G., Assandri, R., Cazzato, D., Rossi, D.S., Usai, S., Tramacere, I., Pellegata, G., Lauria, G., 2020. GI symptoms as early signs of COVID-19 in hospitalised Italian patients. Gut 69, 1547–1548. https://doi.org/10.1136/gutjnl-2020-321434.

Chen, C., Kostakis, C., Gerber, J.P., Tscharke, B.J., Irvine, R.J., White, J.M., 2014. Towards finding a population biomarker for wastewater epidemiology studies. Sci. Total Environ. 487, 621–628. https://doi.org/10.1016/j.scitotenv.2013.11.075.

D'Aoust, P.M., Graber, T.E., Mercier, E., Montpetit, D., Alexandrov, I., Neault, N., Baig, A.T., Mayne, J., Zhang, X., Alain, T., Servos, M.R., 2021. Catching a resurgence: increase in SARS-CoV-2 viral RNA identified in wastewater 48h before COVID-19 clinical tests and 96h before hospitalizations. Sci. Total Environ. 770, 145319 https://doi.org/10.1016/j.scitotenv.2021.145319.

Dejus, B., Cacivkins, P., Gudra, D., Dejus, S., Ustinova, M., Roga, A., Strods, M., Kibilds, J., Boikmanis, G., Ortlova, K., Krivko, L., 2023. Wastewater-based prediction of COVID-19 cases using a random forest algorithm with strain prevalence data: a case study of five municipalities in Latvia. Sci. Total Environ., 164519

El Soufi, G., Di Jorio, L., Gerber, Z., Cluzel, N., Van Assche, J., Delafoy, D., Olaso, R., Daviaud, C., Loustau, T., Schwartz, C., Trebouet, D., Hernalsteens, O., Marechal, V., Raffestin, S., Rousset, D., Van Lint, C., Deleuze, J.F., Boni, M., OBEPINE, consortium, Rohr, O., Wallet, C., et al., 2024. Highly efficient and sensitive membrane-based concentration process allows quantification, surveillance, and sequencing of viruses in large volumes of wastewater. Water Res. 249, 120959 https://doi.org/10.1016/j.watres.2023.120959±.

Farkas, K., Adriaenssens, E.M., Walker, D.I., McDonald, J.E., Malham, S.K., Jones, D.L., 2019. Critical evaluation of CrAssphage as a molecular marker for human-derived wastewater contamination in the aquatic environment. Food Environ. Virol. 11, 113–119. https://doi.org/10.1007/s12560-019-09369-1.

Farkas, K., Hillary, L.S., Thorpe, J., Walker, D.I., Lowther, J.A., McDonald, J.E., Malham, S.K., Jones, D.L., 2021. Concentration and quantification of SARS-CoV-2 RNA in wastewater using polyethylene glycol-based concentration and qRT-PCR. Methods Protoc. 4, 17. https://doi.org/10.3390/mps4010017.

Farkas, K., Pântea, I., Woodhall, N., Williams, D., Lambert-Slosarska, K., Williams, R.C., Grimsley, J.M.S., Singer, A.C., Jones, D.L., 2023. Diurnal changes in pathogenic and indicator virus concentrations in wastewater. Environ. Sci. Pollut. Res. Int. 30, 123785–123795. https://doi.org/10.1007/s11356-023-30381-3.

Farkas, K., Pellett, C., Alex-Sanders, N., Bridgman, M.T.P., Corbishley, A., Grimsley, J.M.S., Kasprzyk-Hordern, B., Kevill, J.L., Pântea, I., Richardson-O'Neill, I.S., Lambert-Slosarska, K., Woodhall, N., Jones, D.L., 2022. Comparative assessment of filtration- and precipitation-based methods for the concentration of SARS-CoV-2 and other viruses from wastewater. Microbiol. Spectr. 10 https://doi.org/10.1128/SPECTRUM.01102-22.

Fedorenko, A., Grinberg, M., Orevi, T., Kashtan, N., 2020. Survival of the enveloped bacteriophage Phi6 (a surrogate for SARS-CoV-2) in evaporated saliva microdroplets deposited on glass surfaces. Sci. Rep. 10, 22419 https://doi.org/10.1038/s41598-020-79625-z.

Feng, S., Roguet, A., McClary-Gutierrez, J.S., Newton, R.J., Kloczko, N., Meiman, J.G., McLellan, S.L., 2021. 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.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 22.

Gendron, L., Verreault, D., Veillette, M., Moineau, S., Duchaine, C., 2010. Evaluation of filters for the sampling and quantification of RNA phage aerosols. Aerosol Sci. Technol. 44, 893–901. https://doi.org/10.1080/02786826.2010.501351.

Gerrity, D., Papp, K., Stoker, M., Sims, A., Frehner, W., 2021. Early-pandemic wastewater surveillance of SARS-CoV-2 in Southern Nevada: methodology, occurrence, and incidence/prevalence considerations. Water Res. X 10, 100086. https://doi.org/10.1016/j.wroa.2020.100086.

Gibas, C., Lambirth, K., Mittal, N., Juel, M.A.I., Barua, V.B., Brazell, L.R., Hinton, K., Lontai, J., Stark, N., Young, I., Quach, C., 2021. . Implementing building-level SARS-CoV-2 wastewater surveillance on a university campus. Sci. Total Environ. 782, 146749 https://doi.org/10.1016/j.scitotenv.2023.168998.

Greenwald, H.D., Kennedy, L.C., Hinkle, A., Whitney, O.N., Fan, V.B., Crits-Christoph, A., Harris-Lovett, S., Flamholz, A.I., Al-Shayeb, B., Liao, L.D., Beyers, M., Brown, D., Chakrabarti, A.R., Dow, J., Frost, D., Koekemoer, M., Lynch, C., Sarkar, P., White, E., Kantor, R., Nelson, K.L., et al., 2021. Tools for interpretation of wastewater SARS-CoV-2 temporal and spatial trends demonstrated with data collected in the San Francisco Bay Area. Water Res. X 12, 100111. https://doi.org/10.1016/j.wroa.2021.100111.

Grolemund, G., Wickham, H., 2011. Dates and times made easy with lubridate. J. Stat. Softw. 40, 1–25. https://doi.org/10.18637/jss.v040.i03.

Haramoto, E., Kitajima, M., Hata, A., Torrey, J.R., Masago, Y., Sano, D., Katayama, H., 2018. A review on recent progress in the detection methods and prevalence of human enteric viruses in water. Water Res. 135, 168–186. https://doi.org/10.1016/j.watres.2018.02.004.

Henriques, T.B., Cassini, S.T., de Pinho Keller, R., 2023. Contribution of wastewater-based epidemiology to SARS-CoV-2 screening in Brazil and the United States. J. Water Health 21, 343–353. https://doi.org/10.2166/wh.2023.260.

Henry, L., Wickham, H., 2020. purrr: Functional Programming Tools. R package version 0.3.4. ⟨https://CRAN.R-project.org/package=purrr⟩.

Hillary, L.S., Farkas, K., Maher, K.H., Lucaci, A., Thorpe, J., Distaso, M.A., Gaze, W.H., Paterson, S., Burke, T., Connor, T.R., McDonald, J.E., Malham, S.K., Jones, D.L., 2021. Monitoring SARS-CoV-2 in municipal wastewater to evaluate the success of lockdown measures for controlling COVID-19 in the UK. Water Res., 117214 https://doi.org/10.1016/j.watres.2021.117214.

Hoffmann, T., Alsing, J., 2023. Faecal shedding models for SARS-CoV-2 RNA amongst hospitalised patients and implications for wastewater-based epidemiology. J. R. Stat. Soc. Ser. C: Appl. Stat. 72, 330–345. https://doi.org/10.1093/jrsssc/qlad011.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 65–70. ⟨http://www.jstor.org/stable/4615733⟩.

Hsu, S.Y., Bayati, M., Li, C., Hsieh, H.Y., Belenchia, A., Klutts, J., Zemmer, S.A., Reynolds, M., Semkiw, E., Johnson, H.Y., Foley, T., Wieberg, C.G., Wenzel, J., Johnson, M.C., Lin, C.H., 2022. Biomarkers selection for population normalization in SARS-CoV-2 wastewater-based epidemiology. Water Res. 223, 118985 https://doi.org/10.1016/J.WATRES.2022.118985.

Jiang, G., Liu, Y., Tang, S., Kitajima, M., Haramoto, E., Arora, S., Choi, P., Jackson, G., D'Aoust, P.M., Delatolla, R., Zhang, S., 2023. Moving forward with COVID-19: future research prospects of wastewater-based epidemiology methodologies and applications. Curr. Opin. Environ. Sci. Health 33, 100458. https://doi.org/10.1016/j.coesh.2023.100458.

Jones, D.L., Baluja, M.Q., Graham, D.W., Corbishley, A., McDonald, J.E., Malham, S.K., Hillary, L.S., Connor, T.R., Gaze, W.H., Moura, I.B., Wilcox, M.H., 2020. Shedding of SARS-CoV-2 in feces and urine and its potential role in person-to-person transmission and the environment-based spread of COVID-19. Sci. Total Environ. 749, 141364 https://doi.org/10.1016/j.scitotenv.2020.141364.

Jung, Y., 2018. Multiple predicting K-fold cross-validation for model selection. J. Nonparametr. Stat. 30, 197–215. https://doi.org/10.1080/10485252.2017.1404598.

Karthikeyan, S., Ronquillo, N., Belda-Ferre, P., Alvarado, D., Javidi, T., Longhurst, C.A., Knight, R., 2021. High-throughput wastewater SARS-CoV-2 detection enables forecasting of community infection dynamics in San Diego County. mSystems 6, e00045-21. https://doi.org/10.1128/mSystems.00045-21.

Kasprzyk-Hordern, B., Sims, N., Farkas, K., Jagadeesan, K., Proctor, K., Wade, M.J., Jones, D.L., 2023. Wastewater-based epidemiology for comprehensive community health diagnostics in a national surveillance study: mining biochemical markers in wastewater. J. Hazard. Mater. 450, 130989 https://doi.org/10.1016/J.JHAZMAT.2023.130989.

Kassambara, A., 2020. ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. ⟨https://CRAN.R-project.org/package=ggpubr⟩.

Kevill, J.L., Pellett, C., Farkas, K., Brown, M.R., Bassano, I., Denise, H., McDonald, J.E., Malham, S.K., Porter, J., Warren, J., Evens, N.P., Paterson, S., Singer, A.C., Jones, D.L., 2022. A comparison of precipitation and filtration-based SARS-CoV-2 recovery methods and the influence of temperature, turbidity, and surfactant load in urban wastewater. Sci. Total Environ. 808, 151916 https://doi.org/10.1016/j.scitotenv.2021.151916.

Kitajima, M., Sassi, H.P., Torrey, J.R., 2018. Pepper mild mottle virus as a water quality indicator. npj Clean. Water 1, 19. https://doi.org/10.1038/s41545-018-0019-5.

Kumar, M., Joshi, M., Patel, A.K., Joshi, C.G., 2021. Unravelling the early warning capability of wastewater surveillance for COVID-19: a temporal study on SARS-CoV-2 RNA detection and need for the escalation. Environ. Res. 196, 110946 https://doi.org/10.1016/j.envres.2021.110946.

Langeveld, J., Schilperoort, R., Heijnen, L., Elsinga, G., Schapendonk, C.E.M., Fanoy, E., de Schepper, E.I.T., Koopmans, M.P.G., de Graaf, M., Medema, G., 2023. 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.

Larremore, D.B., Wilder, B., Lester, E., Shehata, S., Burke, J.M., Hay, J.A., Tambe, M., Mina, M.J., Parker, R., 2021. Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. Sci. Adv. 7, eabd5393 https://doi.org/10.1126/sciadv.abd5393.

Li, X., Liu, H., Gao, L., Sherchan, S.P., Zhou, T., Khan, S.J., Van Loosdrecht, M.C., Wang, Q., 2023. Wastewater-based epidemiology predicts COVID-19-induced weekly new hospital admissions in over 150 USA counties. Nat. Commun. 14 (1), 4548.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22.

Maal-Bared, R., Qiu, Y., Li, Q., Gao, T., Hrudey, S.E., Bhavanam, S., Ruecker, N.J., Ellehoj, E., Lee, B.E., Pang, X., 2023. 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.

McGowan, C.R., Hellman, N., Chowdhury, S., Mannan, A., Newell, K., Cummings, R., 2020. COVID-19 testing acceptability and uptake amongst the Rohingya and host community in Camp 21, Teknaf, Bangladesh. Confl. Health 14, 74. https://doi.org/10.1186/s13031-020-00322-9.
Mitranescu, A., Uchaikina, A., Kau, A.S., Stange, C., Ho, J., Tiehm, A., Wurzbacher, C., Drewes, J.E., 2022. Wastewater-based epidemiology for SARS-CoV-2 biomarkers: evaluation of normalization methods in small and large communities in Southern Germany. ACS ES T Water 2, 2460–2470. https://doi.org/10.1021/acsestwater.2c00306.
Mulvaney, R.L., 1996. Nitrogen – inorganic forms. In: Sparks, D.L., Page, A.L., Helmke, P.A., Loeppert, R.H., Soltanpour, P.N., Tabatabai, M.A., Johnston, C.T., Sumner, M.E. (eds.), Methods of Soil Analysis: Part 3 Chemical Methods. Soil Science Society of America, Madison, WI. ⟨https://doi.org/10.2136/sssabookser5.3.c38⟩.
Murphy, J., Riley, J.P., 1962. A modified single solution methods for the determination of available phosphate in natural water. Anal. Chim. Acta 27, 31–36. https://doi.org/10.1016/S0003-2670(00)88444-5.
Natarajan, A., Zlitni, S., Brooks, E.F., Vance, S.E., Dahlen, A., Hedlin, H., Park, R.M., Han, A., Schmidtke, D.T., Verma, R., Jacobson, K.B., 2022. Gastrointestinal symptoms and fecal shedding of SARS-CoV-2 RNA suggest prolonged gastrointestinal infection. Med 3, 371–387. https://doi.org/10.1016/j.medj.2022.04.001.
ONS, 2021. Census 2021 Geographies. Office for National Statistics, Newport, UK.
Parkins, M.D., Lee, B.E., Acosta, N., Bautista, M., Hubert, C.R.J., Hrudey, S.E., Frankowski, K., Pang, X.L., 2023. Wastewater-based surveillance as a tool for public health action: SARS-CoV-2 and beyond. Clin. Microbiol. Rev., e0010322 https://doi.org/10.1128/cmr.00103-22 (Advance online publication.
Pedersen T.L., 2020. patchwork: The Composer of Plots. R package version 1.1.1. ⟨https://CRAN.R-project.org/package=patchwork⟩.
Puhach, O., Meyer, B., Eckerle, I., 2023. SARS-CoV-2 viral load and shedding kinetics. Nat. Rev. Microbiol. 21, 147–161. https://doi.org/10.1038/s41579-022-00822-w.
R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Rainey, A.L., Liang, S., Bisesi, J.H., Jr, Sabo-Attwood, T., Maurelli, A.T., 2023. A multistate assessment of population normalization factors for wastewater-based epidemiology of COVID-19. PLoS One 18, e0284370. https://doi.org/10.1371/journal.pone.0284370.
Shah, S., Gwee, S.X.W., Ng, J.Q.X., Lau, N., Koh, J., Pang, J., 2022. Wastewater surveillance to infer COVID-19 transmission: a systematic review. Sci. Total Environ. 804, 150060 https://doi.org/10.1016/j.scitotenv.2021.150060.
Shrestha, S., Malla, B., Haramoto, E., 2024. Estimation of Norovirus infections in Japan: an application of wastewater-based epidemiology for enteric disease assessment. Sci. Total Environ. 912, 169334 https://doi.org/10.1016/j.scitotenv.2023.169334.
Shrestha, S., Yoshinaga, E., Chapagain, S.K., Mohan, G., Gasparatos, A., Fukushi, K., 2021. Wastewater-based epidemiology for cost-effective mass surveillance of COVID-19 in low-and middle-income countries: challenges and opportunities. Water 13, 2897. https://doi.org/10.3390/w13202897.
Stachler, E., Kelty, C., Sivaganesan, M., Li, X., Bibby, K., Shanks, O.C., 2017. Quantitative CrAssphage PCR assays for human fecal pollution measurement. Environ. Sci. Technol. 51, 9146–9154. https://doi.org/10.1021/acs.est.7b02703.
Sweetapple, C., Wade, M.J., Melville-Shreeve, P., Chen, A.S., Lilley, C., Irving, J., Grimsley, J.M., Bunce, J.T., 2023. Dynamic population normalisation in wastewater-based epidemiology for improved understanding of the SARS-CoV-2 prevalence: a multi-site study. J. Water Health 21 (5), 625–642.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc.: Ser. B (Methodol.) 58 (1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., Zhao, S., 2019. Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. 18, 463–477. https://doi.org/10.1038/s41573-019-0024-5.
Van Nuijs, A.L., Mougel, J.F., Tarcomnicu, I., Bervoets, L., Blust, R., Jorens, P.G., Neels, H., Covaci, A., 2011. Sewage epidemiology – a real-time approach to estimate the consumption of illicit drugs in Brussels, Belgium. Environ. Int. 37, 612–621. https://doi.org/10.1016/j.envint.2010.12.006.
Vaughan, D., 2020. slider: Sliding Window Functions. R package version 0.1.5. ⟨https://CRAN.R-project.org/package=slider⟩.
Wade, M.J., Jacomo, A.L., Armenise, E., Brown, M.R., Bunce, J.T., Cameron, G.J., Fang, Z., Farkas, K., Gilpin, D.F., Graham, D.W., Grimsley, J.M., et al., 2022. 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.
Wang, C., Mao, G., Liao, K., Ben, W., Qiao, M., Bai, Y., Qu, J., 2021. Machine learning approach identifies water sample source based on microbial abundance. Water Res. 199, 117185 https://doi.org/10.1016/j.watres.2021.117185.
Wannigama, D.L., Amarasiri, M., Phattharapornjaroen, P., Hurst, C., Modchang, C., Chadsuthi, S., Anupong, S., Miyanaga, K., Cui, L., Thuptimdang, W., Ali Hosseini Rad, S.M., Fernandez, S., Huang, A.T., Vatanaprasan, P., Jay, D.J., Saethang, T., Luk-In, S., Storer, R.J., Ounjai, P., Ragupathi, N.K.D., Hongsing, P., et al., 2023. Tracing the transmission of mpox through wastewater surveillance in Southeast Asia. J. Travel Med. 30, taad096 https://doi.org/10.1093/jtm/taad096.
Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
Wickham, H., 2020. tidyr: Tidy Messy Data. R package version 1.1.2. ⟨https://CRAN.R-project.org/package=tidyr⟩.
Wickham, H., François, R., Henry, L., Müller, K., 2021. dplyr: A Grammar of Data Manipulation. R package version 1.0.4. ⟨https://CRAN.R-project.org/package=dplyr⟩.
Wilder, M.L., Middleton, F., Larsen, D.A., Du, Q., Fenty, A., Zeng, T., Insaf, T., Kilaru, P., Collins, M., Kmush, B., Green, H.C., 2021. Co-quantification of crAssphage increases confidence in wastewater-based epidemiology for SARS-CoV-2 in low prevalence areas. Water Res. X 11, 100100. https://doi.org/10.1016/j.wroa.2021.100100.
Wu, F., Zhang, J., Xiao, A., Gu, X., Lee, W.L., Armas, F., Kauffman, K., Hanage, W., Matus, M., Ghaeli, N., Endo, N., Duvallet, C., Poyet, M., Moniz, K., Washburne, A. D., Erickson, T.B., Chai, P.R., Thompson, J., Alm, E.J., 2020. SARS-CoV-2 titers in wastewater are higher than expected from clinically confirmed cases. mSystems 5, e00614-20. https://doi.org/10.1128/mSystems.00614-20.
Zakhour, M., Maalouf, H., Di Bartolo, I., Haugarreau, L., Le Guyader, F.S., Ruvoën-Clouet, N., Le Saux, J.C., Ruggeri, F.M., Pommepuy, M., Le Pendu, J., 2010. Bovine norovirus: carbohydrate ligand, environmental contamination, and potential cross-species transmission via oysters. Appl. Environ. Microbiol. 76, 6404–6411. https://doi.org/10.1128/AEM.00671-10.
Zeileis, A., Grothendieck, G., 2005. zoo: S3 infrastructure for regular and irregular time series. J. Stat. Softw. 14, 1–27. https://doi.org/10.18637/jss.v014.i06.
Zhang, H., Nettleton, D., Zhu, Z., 2019. Regression-enhanced random forests. arXiv preprint arXiv:1904.10416. ⟨https://doi.org/10.48550/arXiv.1904.10416⟩.
Zhang, S., Shi, J., Li, X., Tiwari, A., Gao, S., Zhou, X., Sun, X., O'Brien, J.W., Coin, L., Hai, F., Jiang, G., 2023. Wastewater-based epidemiology of *Campylobacter* spp.: a systematic review and meta-analysis of influent, effluent, and removal of wastewater treatment plants. Sci. Total Environ. 903, 166410 https://doi.org/10.1016/j.scitotenv.2023.166410.
Zhu, Y., Oishi, W., Maruo, C., Saito, M., Chen, R., Kitajima, M., Sano, D., 2021a. Early warning of COVID-19 via wastewater-based epidemiology: potential and bottlenecks. Sci. Total Environ. 767, 145124 https://doi.org/10.1016/j.scitotenv.2021.145124.
Zhu, Y., Oishi, W., Saito, M., Kitajima, M., Sano, D., 2021b. Early warning of COVID-19 in Tokyo via wastewater-based epidemiology: how feasible it really is? J. Water Environ. Technol. 19, 170–183. https://doi.org/10.2965/jwet.21-024.