Monitoring SARS-CoV-2 Using Infoveillance, National Reporting Data and Wastewater in Wales, UK

Jordan P. Cuff[1,2+], Shrinivas Nivrutti Dighe[1], Sophie E. Watson[1], Rafael A. Badell-Grau[3], Andrew J. Weightman[1+], Davey L. Jones[4], Peter Kille[1+]

[1]School of Biosciences, Cardiff University, Sir Martin Evans Building, Museum Avenue, Cardiff, CF10 3AX, UK

[2]School of Natural and Environmental Sciences, Newcastle University, Newcastle-upon-Tyne, NE1 7RU, UK

[3]Division of Genetics, Department of Paediatrics, University of California, San Diego, La Jolla, CA, 92093, USA

[4]School of Natural Sciences, Bangor University, Bangor, Gwynedd, LL57 2UW, UK

[+]Corresponding authors: jordancuff@gmail.com, Weightman@cardiff.ac.uk, Kille@cardiff.ac.uk

Abstract

Background: The COVID-19 pandemic necessitated rapid real-time surveillance of epidemiological data to advise governments and the public, but the accuracy of these data depend on myriad auxiliary assumptions, not least accurate reporting of cases by the public. Wastewater monitoring has emerged internationally as an accurate and objective means for assessing disease prevalence with reduced latency and less dependence on public vigilance, reliability, and engagement. How public interest aligns with COVID-19 personal testing data and wastewater monitoring is, however, very poorly characterised.

Objectives: This study assesses the associations between internet search volume data relevant to COVID-19, public healthcare statistics and national-scale wastewater monitoring of SARS-CoV-2 across South Wales, UK over time to investigate how interest in the pandemic may reflect the prevalence of SARS-CoV-2, as detected by national testing and wastewater monitoring, and how these data could be used to predict case numbers.

Methods: Relative search volume data from Google Trends for search terms linked to the COVID-19 pandemic were extracted and compared against government-reported COVID-19 statistics and RT-qPCR SARS-CoV-2 data generated from wastewater in South Wales, UK, using multivariate linear models, correlation analysis and predictions from linear models.

Results: Wastewater monitoring and infoveillance both show potential for epidemiological surveillance, but their efficacy changes over time. Google search volumes surrounding the COVID-19 pandemic decreased across the study period, suggesting a reduction in public interest which may be reflected in lower volumes of self-testing and reporting with subsequent loss of accuracy of national reporting data.

Conclusions: Wastewater monitoring presents a valuable means for assessing population-level prevalence of SARS-CoV-2 and could be integrated with other data types such as infoveillance for increasingly accurate inference of virus prevalence. The importance of such monitoring is increasingly clear as a means of objectively assessing the prevalence of SARS-CoV-2 to circumvent the dynamic interest and participation of the public. Increased accessibility of wastewater monitoring data to the public, as is the case for other national data, may enhance public engagement with these forms of monitoring.

Keywords: COVID-19, Google Trends, Infodemiology, RT-qPCR, Wastewater

Introduction

The COVID-19 pandemic has given rise to a range of public responses which have dynamically driven cooperation of the public with governmental guidance, and public recognition of the need for regular testing. Healthcare systems have been stretched beyond capacity by sudden, large-volume influxes of patients following sometimes unpredictable waves of the virus [1]. There is a pressing need for local, national and global adaptability to manage these outbreaks of the disease to minimise the impact on healthcare systems, the first requirement of which is stringent collection of reliable and accurate data of viral prevalence [2].

Many strategies have been employed to monitor SARS-CoV-2, including self-reporting [3] and participatory surveillance [4–6], including through use of platforms such as accessible phone apps [7]. Surveys and self-reporting, achieved through participatory surveillance and even active crowdsourcing strategies, have proven highly effective through monitoring of symptoms such as loss-of-taste [8]; participatory surveillance platforms such as this have been a crucial component of monitoring in partnership with the public [8,9]. Relying on surveys and personal testing data, however, allows only a reactive approach to mitigating the healthcare burden imposed by COVID-19, which is often too little, too late to mitigate the heavy case numbers and death tolls. Case data, while sometimes collected by standardised surveys, can otherwise depend on self-reporting by the public, many members of which may not self-test given poor access to tests, may not feel obliged due to asymptomatic cases or may receive false negative results. Others may unreliably or even dishonestly report the results of tests given the restrictions that a positive test for COVID-19 imposed [10], or they may be disenfranchised with the efforts to reduce the prevalence of the disease given the overwhelming extent of misinformation in circulation [11].

Search engine usage has been explored as a means for ascertaining the prevalence of diseases [12,13], but this method is not infallible and its accuracy over time must be assessed in different epidemiological contexts [14,15]. Such data can, however, track COVID-19 or specific related symptoms [16–19] given its broad accessibility. Whilst searching for particular character strings cannot be reliably and accurately associated with the prevalence of the disease, this ‘infoveillance’ does facilitate analysis of public interest in subjects such as the pandemic [11,20], which can be an important factor in healthcare management and the pandemic response. Infoveillance can be integrated into interdisciplinary frameworks such as ‘One Health’ [21,22] and, more specifically, ‘One Digital

Health' [23], which aim to view healthcare matters more holistically, particularly the interaction between human and veterinary health and its implications for zoonotic diseases, but also the environmental dimension of disease occurrence and transmission.

Given the latency of surveys and testing by the public, and the potential inaccuracies of infoveillance approaches, objective means for disease surveillance without the requirement of public participation have become increasingly important throughout the COVID-19 pandemic. The presence of coronaviruses and other human pathogenic viruses in human faeces and their subsequent presence in urban wastewater is a long-established tool for assessing disease prevalence within a community [24,25]. This approach provides a non-invasive means for assessing SARS-CoV-2 prevalence across whole populations via wastewater [25–31]. Monitoring of wastewater has provided a robust and accurate means of assessing the population-level prevalence of COVID-19, facilitating some prediction of healthcare burden before symptoms arise [32]. Wastewater monitoring circumvents several barriers preclusive to accurate testing data such as hesitancy, availability of testing, asymptomatic patients and socio-economic or cultural barriers by passively sampling from whole communities [10,33]. The efficacy of this approach does not depend on public participation, possibly leading some inconsistencies with national testing statistics. A strong positive correlation between direct testing, wastewater monitoring data and public interest in the pandemic has been demonstrated [34] but the dynamic relationship between these data and how public interest dictates the accuracy of monitoring data are still poorly characterised.

Here, we compare public interest in the pandemic through search engine usage data against wastewater SARS-CoV-2 surveillance data and nationally reported statistics over time to assess how public interest dictated the relationship between disease prevalence and reporting over a year of the COVID-19 pandemic in South Wales, UK. This study also explores the efficacy of wastewater monitoring and infoveillance as means for assessing the national state of the pandemic, how these relationships change over time and how they could inform predictions of case numbers for streamlined monitoring.

Methods

Wastewater monitoring

Since mid-September 2020, wastewater samples were collected every Monday, Wednesday, and Friday from Cardiff Bay, Newport Nash, Llanfoist, Ponthir, Ogmore, Cog Moors, Swansea Bay, and Gowerton wastewater treatment plants, and samples from Carmarthen and Haverfordwest were collected every Wednesday. Samples were transported on ice in a cooler box to designated wastewater processing facilities at Cardiff University. Processing of samples was based on Farkas *et al.* [35]. From each site, 200 ml of wastewater was spun at 3000 x *g* for 30 min, and 150 ml of supernatant was neutralised to pH 7-7.4 using 1 M NaOH. The supernatant was incubated with 50 ml of 40% PEG and 8% NaCl overnight. Samples were then spun at 10,000 x *g* for 30 min and the pellet was dissolved in 500 µl of PBS (pH-7.4). Of the dissolved pellet, 100 µl was spiked with 10,000 copies of synthetic Murine Norovirus (MNV) DNA to check the extraction efficiency. Subsequent nucleic acid extraction and amplification took place in the COVID-19 testing facilities at Cardiff University. Total RNA was extracted using the methodology published by Oberacker et al. [36]. Total RNA was eluted in 100 µl of nuclease-free water. For SARS-CoV-2 detection, four primer sets published by the United States Centers for Disease Control and Prevention (US CDC), Charité and Hong Kong University [37] were used for RT-qPCR. Primer sets N1 and N2 target different regions of nucleocapsid (N genes); E_Sarbeco and ORF1b target the SARS-CoV-2 E and nsp14 genes, respectively. For the controls, a set of primers that target virus crAssphage [38] (which is present in human faecal material) and MNV [39] (which was used to assess extraction efficiency) were selected (Table 1). Samples were run in triplicate on Fast 384-well plates (Applied Biosystems) using QuantStudio 7 Flex (Applied Biosystems). A 10 µl RT-qPCR reaction was performed containing 4 µl of extracted RNA template, 5 µl of Luna Universal Probe One-step Reaction Mix (2X; NEB), 0.04 µl of each primer set (100 µM), 0.02 µl of fluorescent probe (100 µM), 0.5 µl NEB Luna reverse transcriptase (20X) and 0.4 µl nuclease free water. The reverse transcription (RT) was carried out at 55 °C for 10 min, followed by polymerase activation at 95.0 °C for 1 min, and 40 cycles of denaturation, annealing and extension at 95.0 °C for 10 sec, and then 60.0 °C for 1 min, respectively. Serial dilutions of the heat-inactivated SARS-CoV-2 viral standards were run on every PCR plate to generate standard curves used to quantify the copies of SARS-CoV-2 genes. Additionally, RT-qPCR runs were validated by positive (Qnostics, SCV2QC01-QC) and negative controls (nuclease-free water). Resultant data were normalised to account for population size in each area, and to correct for dilution as described by Wilde et al. [40].

Table 1: The qPCR primers used for wastewater monitoring.

Assay	Target gene	Sequences (5’--3’)
E_Sarbeco	E	F- 5’-ACAGGTACGTTAATAGTTAATAGCGT-3’ R- 5’-ATATTGCAGCAGTACGCACACA-3’ P- 5’-HEX-ACACTAGCCATCCTTACTGCGCTTCG- IB®FQ-3
ORF1b	nsp14	F-5’-TGGGGYTTTACRGGTAACCT-3’ R-5’-AACRCGCTTAACAAAGCACTC-3’ P- 5’-FAM-TAGTTGTGATGCWATCATGACTAG- IB®FQ-3’
N1	Nucleocapsid	F- 5’-GACCCCAAAATCAGCGAAAT-3’ R- 5’-TCTGGTTACTGCCAGTTGAATCTG-3’ P-5-HEX- ACCCCGCATTACGTTTGGTGGACC- IB®FQ-3’
N2	Nucleocapsid	F- 5’-TTACAAACATTGGCCGCAAA-3’ R- 5’-GCGCGACATTCCGAAGAA-3’ P- 5’-FAM- ACAATTTGCCCCCAGCGCTTCAG-IB®FQ-3’
crAssphage	Q56	F- 5’-CAGAAGTACAAACTCCTAAAAAACGTAGAG-3’ R- 5’-GATGACCAATAAACAAGCCATTAGC-3’ P- 5’-HEX- AATAACGATTTACGTGATGTAAC- IB®FQ-3’
MNV	-	F- 5’-CCGCAGGAACGCTCAGCAG-3’ R- 5’-GGYTGAATGGGGACGGCCTG-3’ P-5’-FAM- ATGAGTGATGGCGCA- IB®FQ-3’

National statistics and search volume data extraction

This study concerns two periods: the primary study period (between the weeks of 2020-10-11 and 2021-10-31; the focus of all analyses and visualisations aside from comparison with model-based predictions described below) and the full study period (the primary study period with extension up to 2022-07-17 to facilitate comparison of real-world data with model-based predictions). All data were generated or extracted to encompass the full study period. National statistics on the daily number of COVID-19 cases, deaths and vaccinations in Wales were extracted from the UK Government COVID-19 data portal for the full study period [41]. Case data were new cases by publish date (i.e., the number of new cases reported since the previous update; API = ‘newCasesByPublishDate’). Death data were new daily national statistics office deaths by death date (i.e., daily numbers of deaths of people whose death certificate mentioned COVID-19 as one of the causes; API = ‘newDailyNsoDeathsByDeathDate’). Vaccine data were new vaccines given by publish date (i.e., daily numbers of new vaccines (all doses) given; API =

‘newVaccinesGivenByPublishDate’). These data can be downloaded via a permanent download link [41].

Search volume data were extracted from Google Trends. These data provide a proxy for public interest in or response to the extent of the COVID-19 pandemic. The data extracted from Google Trends are relative search volumes for pre-determined search terms, allowing comparison of search rates for different terms via Google, the most widely used internet search engine. These relative search volumes are presented for each date of a given period within a given country, nation or region and are normalised relative to the highest search volume peak in that search batch in the time period specified (this peak represented as a search volume of 100 %). Search volumes were releveled so that the highest peak in the primary study period was represented by ‘100’ and any higher peaks across the full study period exceeded 100 to reflect the limitations of making real-time predictions from existing data. Given the representation of numbers less than one as “<1” by Google Trends, all relative search volumes of “<1” were converted to 0 to facilitate quantitative comparison.

Search terms were selected based on their broad relevance throughout the study period and the high volume of searches generated during that period. These included: “COVID lockdown”, “COVID rules”, “COVID symptoms”, “COVID test” and “COVID vaccine”. “COVID” was included in each search term to ensure relevance to the COVID-19 pandemic; “COVID” was selected over “coronavirus”, “SARS-CoV-2” and other variations due to the greater prevalence of searches related to this string, and its inclusion within other search strings like “COVID-19”.

Statistical analysis

Statistical analyses and plotting of data were carried out using R version v4.0.3 [42] and all data and code are openly available [43]. Since wastewater sites were sampled weekly, all data were averaged first by site and then by week. Wastewater qPCR data were log-transformed to improve model fit and visualisation. Data were processed and aggregated using ‘tidyverse’ packages for reproducibility [44].

Correlations between search volumes, wastewater SARS-CoV-2 prevalence and nationally reported cases, deaths and vaccinations were tested using Spearman’s rho rank correlation via the ‘rcor’ function of the ‘Hmisc’ package [45]. To facilitate assessment of correlation, week dates were transformed into successive study weeks (i.e., cumulative weeks of the study). The data were identified as non-normally distributed via Shapiro-Wilk tests, so non-parametric correlation analyses were selected. The output was visualised in a correlogram via the ‘corrplot’ function of the ‘corrplot’ package [46], with colours to denote the strength of correlations assigned via the ‘viridis’ package [47].

To assess how relative search volume for the selected search terms changed with differences in the number of COVID-related cases, deaths and vaccines, and the estimated prevalence of COVID in wastewater, a multivariate linear model (MLM) was built via ‘manylm’ in the ‘mvabund’ package [4]. The dependent variable comprised the relative search volumes, log transformed (log[n+1]) to achieve normality, and the independent variables were week, national cases, deaths and vaccinations, and two-way interactions between study week and each of the other variables. For visualisation via line plots, data were releveled so that their minimum and maximum values were 0 and 100, respectively. These normalised search volume, wastewater and government data were plotted against time using the ‘ggplot2’ package [48], with colours assigned via the ‘paired’ palette in the ‘RColorBrewer’ package [49] and data lines smoothed using the ‘loess’ method.

Pairwise plots were generated for reported case data, qPCR data and RSVs from each of the Google Trends search terms separately using ‘ggpairs’ from the ‘GGAlly’ package. Linear models were generated with the number of reported cases as the dependent variable and, in a separate model for each, the qPCR and Google Trends data as independent variables. The ‘predict’ function was used to make interpolated predictions of case numbers across the primary study period, and extrapolated predictions of case numbers beyond the primary study period for the remainder of the full study period. These predicted case numbers were plotted against the reported case numbers, and a correlation analysis carried out as described above. A generalized linear model with a gaussian error family was built with reported cases as the dependent variables, and predicted case numbers, time and pairwise interactions between predictions and time as independent variables.

Information sources and reliability

Wastewater monitoring data were generated by authors of this study at Cardiff University as part of the Welsh Government-funded WEWASH project. The national statistics on COVID-19 cases, deaths and vaccinations were extracted from the UK Government COVID-19 data portal [41], which is internationally recognised as a reputable source used for national reporting, scientific research and public awareness. The Google Trends data should be reliable as indicators of Google usage since they are collected by Google based on the input of users of their service.

Results

Overall, significant correlations were identified between many of the variables (Figure 1, Table S1). Notably, wastewater SARS-CoV-2 RNA prevalence significantly positively correlated with the number of reported cases (Spearman rho = 0.428, P = .001), but did not correlate with the number of reported deaths (Spearman rho = 0.044, P = .746). Of the search terms included, wastewater prevalence positively correlated with “COVID symptoms” (Spearman rho = 0.369, P = .005), “COVID test” (Spearman rho = 0.356, P = .007), and significantly negatively correlated with “COVID vaccine” (Spearman rho = -0.504, P < .001). The number of reported cases, however, positively correlated with both “COVID symptoms” (Spearman rho = 0.805, P < .001) and “COVID test” (Spearman rho = 0.531, P < .001), but negatively correlated with “COVID vaccine” (Spearman rho = -0.495, *P* = .001). All search terms except “COVID rules” significantly negatively correlated with national vaccinations (Table S1).

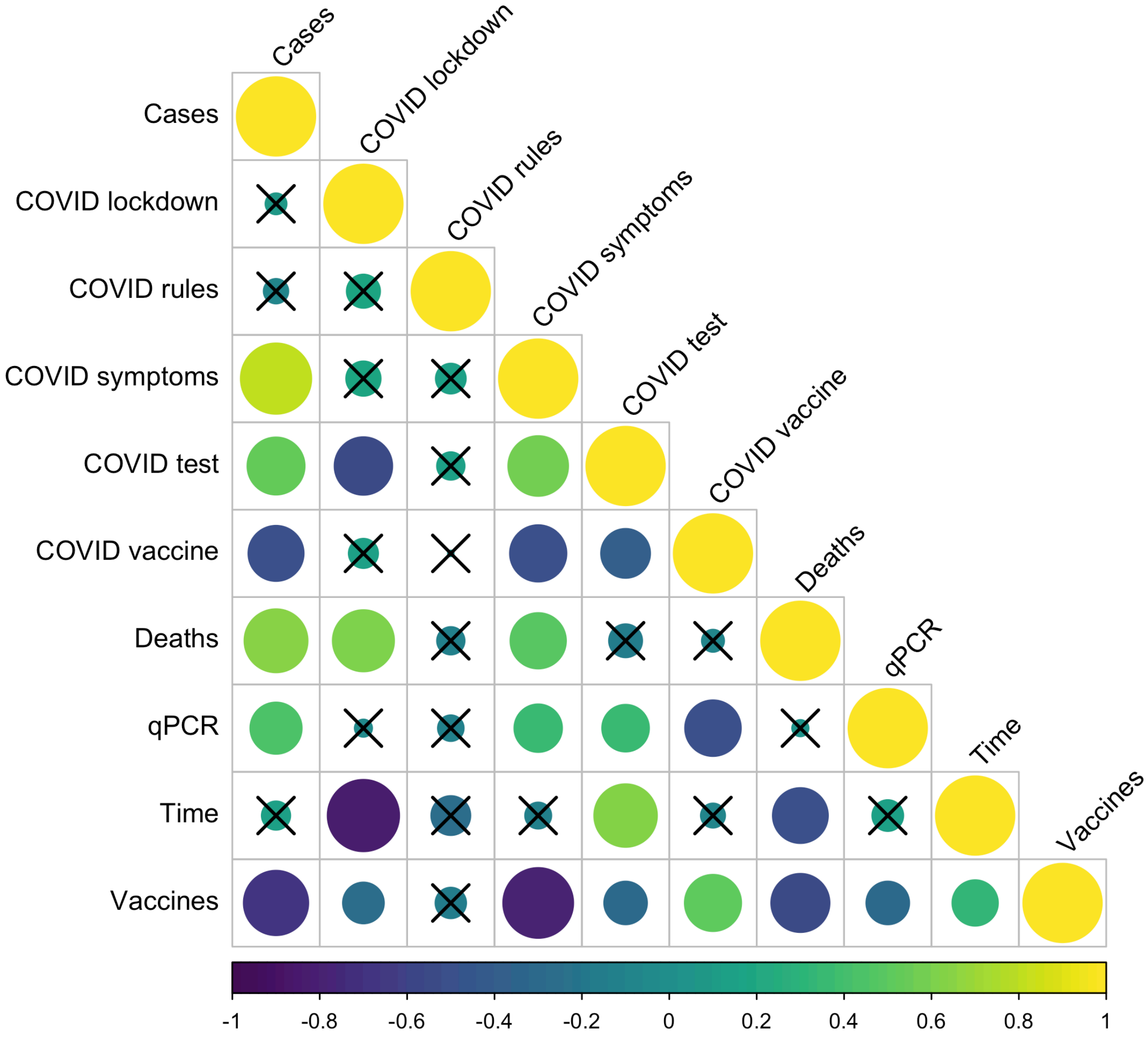

Figure 1. Correlogram of time (study week, i.e., progressive number of weeks into the study period), Google Trends search volumes (variables starting with ‘COVID’), nationally reported cases, deaths and vaccinations, and qPCR-based wastewater SARS-CoV-2 RNA prevalence. Circle size and colour (purple, through teal to yellow - denoting negative through neutral to positive) indicate the extent and directionality of the correlation. Crossed out circles are those for which correlations were not significant.

Search volumes were significantly related to several of the independent variables and their interactions (Table 2; Figure 2), comprising wastewater SARS-CoV-2 prevalence (MLM: $F_{1,54} = 34.89$, $P = .002$), time (MLM: $F_{1,53} = 120.89$, $P = .002$), national COVID-19 cases reported (MLM: $F_{1,52} = 117.77$, $P = .002$), national COVID-19-related deaths reported (MLM: $F_{1,51} = 65.84$, $P = .002$), national COVID-19 vaccines administered (MLM: $F_{1,50} = 54.31$, $P = .002$), and the interactions between time and national COVID-19 cases (MLM: $F_{1,48} = 46.32$, $P = .002$), time and national COVID-19 deaths (MLM: $F_{1,48} = 26.09$, $P = .004$), and time and national vaccinations (MLM: $F_{1,46} = 15.10$, $P = .022$). The interaction between time and wastewater SARS-CoV-2 RNA prevalence (MLM: $F_{1,49} = 0.77$, $P = .967$) was not significantly related to relative search volumes.

Table 2: Univariate results from the multivariate linear model results for search volume data analysed against time (progressive study weeks), wastewater SARS-CoV-2 RNA prevalence, nationally reported COVID-19 cases, deaths and vaccines, and two-way interactions between time and each other variable.

Independent variable	**"COVID symptoms"**	**"COVID test"**	**"COVID vaccine"**	**"COVID rules"**	**"COVID lockdown"**
Wastewater SARS-CoV-2 prevalence	$F_{1,54} = 2.211$, $P = .338$	$F_{1,54} = 0.418$, $P = .687$	$F_{1,54} = 28.838$, $P = .002$	$F_{1,54} = 0.583$, $P = .687$	$F_{1,54} = 2.834$, $P = .306$
Time	$F_{1,53} = 0.189$, $P = .885$	$F_{1,53} = 34.716$, $P = .002$	$F_{1,53} = 0.120$, $P = .885$	$F_{1,53} = 4.414$, $P = .116$	$F_{1,53} = 81.453$, $P = .002$
National COVID-19 cases reported	$F_{1,52} = 77.157$, $P = .002$	$F_{1,52} = 28.501$, $P = .002$	$F_{1,52} = 4.122$, $P = .106$	$F_{1,52} = 0.677$, $P = .413$	$F_{1,52} = 7.315$, $P = .030$
National COVID-19-related deaths	$F_{1,51} = 2.373$, $P = .222$	$F_{1,51} = 13.42$, $P = .003$	$F_{1,51} = 18.621$, $P = .003$	$F_{1,51} = 30.232$, $P = .002$	$F_{1,51} = 1.193$, $P = .245$
Vaccines administered nationally	$F_{1,50} = 17.880$, $P = .002$	$F_{1,50} = 21.308$, $P = .002$	$F_{1,50} = 8.766$, $P = .025$	$F_{1,50} = 0.586$, $P = .429$	$F_{1,50} = 5.770$, $P = .048$
Time: Wastewater prevalence	$F_{1,49} = 0.284$, $P = .982$	$F_{1,49} = 0.067$, $P = .982$	$F_{1,49} = 0.011$, $P = .982$	$F_{1,49} = 0.243$, $P = .982$	$F_{1,49} = 0.165$, $P = .982$
Time: Cases	$F_{1,48} = 3.349$, $P = .157$	$F_{1,48} = 15.165$, $P = .002$	$F_{1,48} = 10.632$, $P = .004$	$F_{1,48} = 15.869$, $P = .002$	$F_{1,48} = 1.301$, $P = .266$
Time: Deaths	$F_{1,47} = 3.536$, $P = .181$	$F_{1,47} = 4.113$, $P = .154$	$F_{1,47} = 3.04$, $P = .181$	$F_{1,47} = 0.246$, $P = .594$	$F_{1,47} = 15.155$, $P =.004$
Time: Vaccines	$F_{1,46} = 0.241$, $P = .813$	$F_{1,46} = 0.171$, $P = .813$	$F_{1,46} = 6.898$, $P = .062$	$F_{1,46} = 1.89$, $P = .367$	$F_{1,46} = 5.903$, $P = .069$

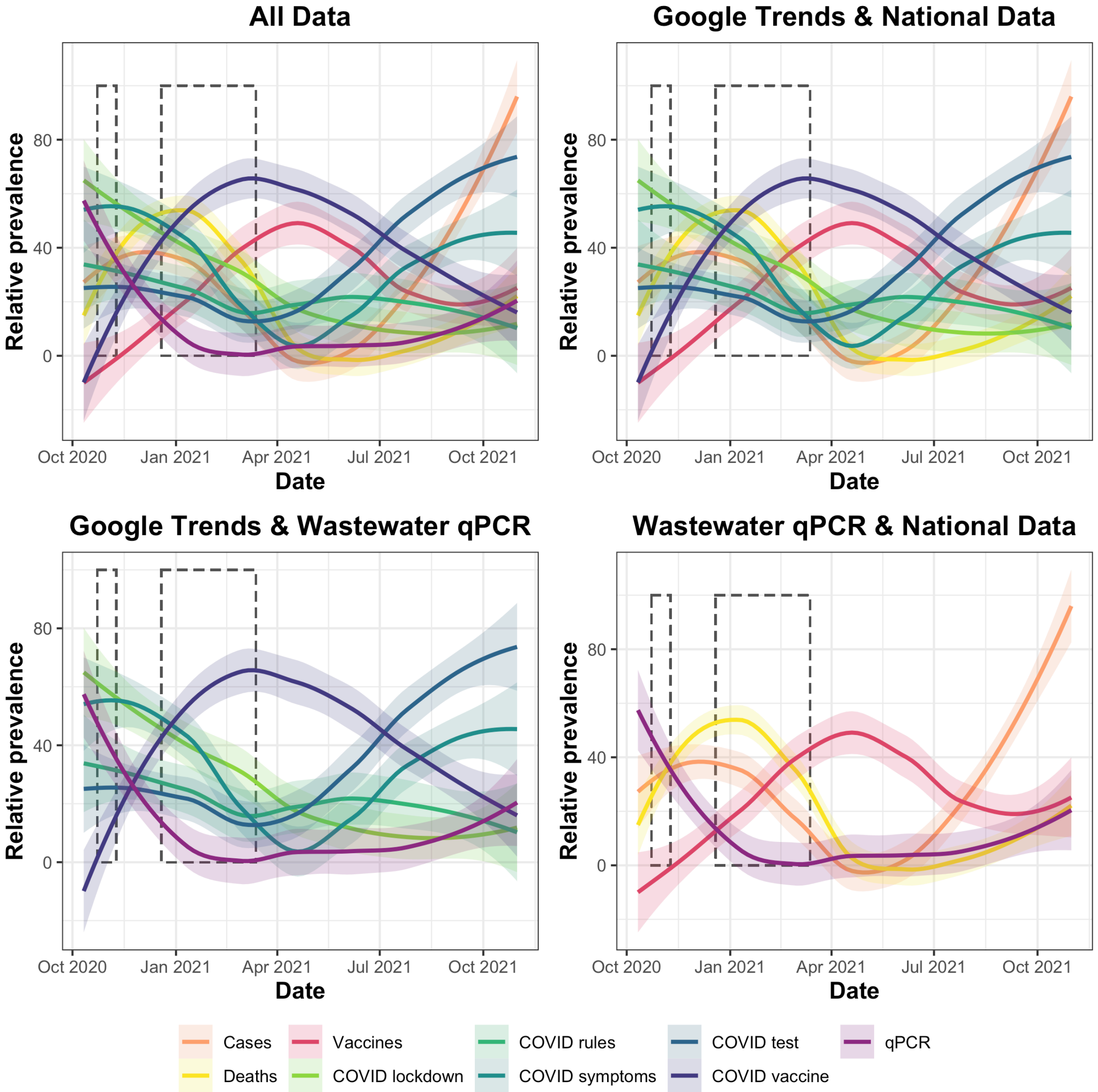

Figure 2. Relative search volumes extracted from Google Trends compared against nationally reported data and qPCR-based estimates of prevalence for SARS-CoV-2 in wastewater. All values are normalised so that the maximum value for each variable is 100. Lines are loess smoothed curves, thus represent the overall trend and do not always represent the most extreme (e.g., maximum) values. Dashed rectangles represent periods of national lockdown in Wales for reference. Wastewater qPCR-based SARS-CoV-2 prevalence is given in light purple, Google Trends data are given in green/blue and national data are given in orange/red/purple. A figure containing non-smoothed trends is presented in Figure S1.

National case data significantly related to Google Trends data for “COVID symptoms” (LM: $t = 7.248$, $P < .001$), “COVID test” (LM: $t = 6.070$, $P < .001$) and “COVID vaccine” (LM: $t = -3.301$, $P = .002$), but not qPCR-based wastewater SARS-CoV-2 prevalence (LM: $t = 1.360$, *P* = .179), nor Google Trends data for “COVID lockdown” (LM: t = 0.897, *P* = .374) and “COVID rules” (LM: t = 0.320, *P* = .750). Significant correlations were identified between case data and some the predicted case numbers (Figure 3, Table S2). Notably, wastewater SARS-CoV-2 RNA prevalence-based predictions significantly positively correlated with the number of reported cases (Spearman rho = 0.274, *P* = .008). Of the search terms included, case data correlated with predictions based on “COVID symptoms” (Spearman rho = 0.683, *P* < .001), “COVID test” (Spearman rho = 0.706, *P* < .001) and “COVID rules” (Spearman rho = 0.409, *P* < .001). National case data significantly related to case numbers predicted by “COVID symptoms” (GLM: t = 5.158, *P* < 0.001) and “COVID test” (GLM: t = -4.997, *P* < 0.001) RSVs, but these relationships changed over time (“COVID symptoms”: t = -5.162, *P* < 0.001; “COVID test”: t = 5.029, *P* < 0.001; Figure 4). National case data marginally insignificantly related to case numbers predicted by qPCR wastewater SARS-CoV-2 prevalence (GLM: t = -1.896, *P* = 0.016) and “COVID rules” RSVs (GLM: t = 1.853, *P* = 0.068), but these relationships were marginally insignificantly related to time (qPCR: t = 1.920, *P* = 0.059 “COVID rules”: t = -1.866, *P* = 0.066; Figure 4).

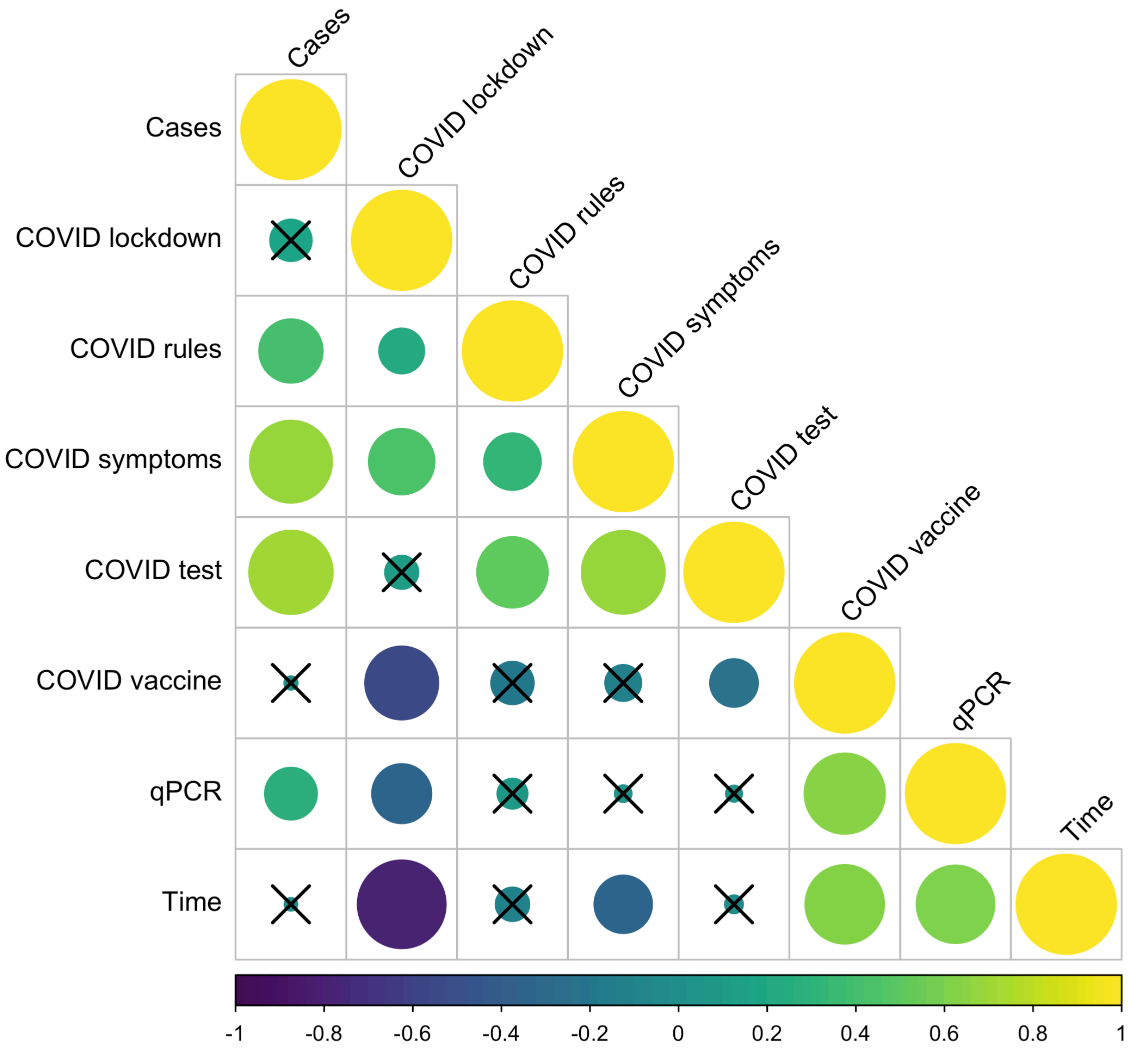

Figure 3. Correlogram of time (study week, i.e., progressive number of weeks into the study period), nationally reported cases, and the number of cases predicted based on linear models of cases against Google Trends search volumes and qPCR-based wastewater SARS-CoV-2 prevalence. Circle size and colour (purple, through teal to yellow - denoting negative through neutral to positive) indicate the extent and directionality of the correlation. Crossed out circles are those for which correlations were not significant.

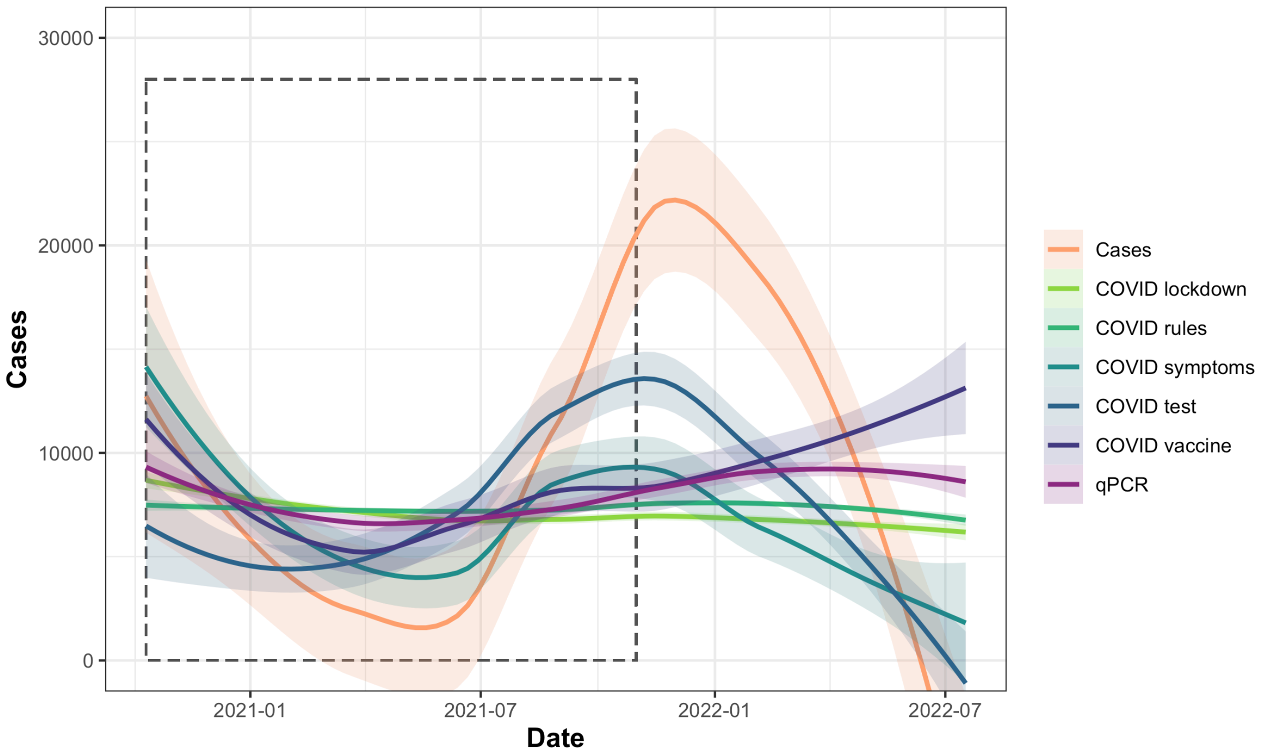

Figure 4. COVID-19 case numbers, and predicted case numbers interpolated and extrapolated based on linear models of case numbers and, separately, each Google Trends search term and qPCR-based SARS-CoV-2 prevalence in wastewater. The dashed rectangle denotes the primary study period, within which data are interpolated. Interpolations are based on data from the primary study period from which models were generated. Extrapolations (outside of the rectangle) are based on data from the following nine months. Wastewater qPCR-estimated SARS-CoV-2 prevalence is given in light purple, Google Trends data are given in green/blue and national reported case data are given in orange. Non-smoothed data are presented in Figure S7.

Discussion

This study provides evidence to suggest that public interest in topics related to the pandemic changed dynamically across the study period, with some relation to the prevalence of the virus in wastewater, and the number of reported cases. Both internet search volume and qPCR-based SARS-CoV-2 RNA prevalence data provide some predictive potential for monitoring of SARS-CoV-2 and could be applied across other contexts.

During the course of this study, comprising two significant waves of the COVID-19 pandemic in Wales, the correlation between reported COVID-19 cases and wastewater-quantified SARS-CoV-2 prevalence was significantly positive overall, as has been demonstrated in previous studies [28,34], but this correlation may have changed over time. Comparing the prevalence of wastewater SARS-CoV-2 estimates and national cases across the full study period shows that wastewater prevalence of SARS-CoV-2 peaked substantially higher in early Autumn 2020 than the rest of the study period, whereas case data peaked the following Autumn (Figure S1). Indications of correlation between SARS-CoV-2 prevalence in wastewater and COVID-19 disease prevalence were recognised at an early stage of the pandemic in other countries [32]. The Google Trends search volume data show online searching for some COVID-19-related strings largely reduced over time, although this was highly dependent on the search string. This could indicate reduced public interest, fluctuations in which were reported even in the initial months of the pandemic despite the importance of sustained public action to ensure the success of public health measures [50]. In this same period, many of the search volumes, with the intuitive exception of 'COVID vaccine', appear to inversely correlate with increased vaccinations. This suggests that the public may have been seeking vaccine opportunities and otherwise expressed less interest in COVID-19 following mass vaccinations, although additional data would be required to confirm this. Importantly, searches for "COVID vaccine" may also represent those that were concerned with misinformation or conspiracy theories which were commonplace, particularly around the vaccine [11].

The search term 'COVID test' was maintained at a relatively constant level throughout the study and, along with 'COVID symptoms' and 'COVID vaccine', correlated with the wastewater SARS-CoV-2 prevalence just as national case data did. This indicates the potential of carefully selected search terms for estimating the prevalence of the virus, further ratified by the predictions made in this study. The relationship between predictions and case data varied greatly depending on the data used to guide predictions and, importantly, these relationships changed over time. The variable potential of infoveillance to predict epidemiological trends has been recorded in other cases, such as for Google Flu Trends [13,15] and is an important consideration for the use of infoveillance in a monitoring context.

The efficacy of infoveillance is contingent on public interest consistently reflecting epidemiology which is ultimately unlikely for global pandemics given natural spikes and fluctuations in public interest. It is, however, important to contextualise this with the likely reasons for members of the public searching with this particular string. Search volume data could nonetheless provide anecdotal monitoring of disease prevalence, especially since many nations face difficulties in monitoring the virus using molecular methods or population-level testing. Search volume data, while imperfect, may provide a valuable alternative for anecdotal epidemiological monitoring in nations or regions lacking access to alternatives [51], but the search terms must be carefully considered, closely monitored and interpreted with appropriate scepticism.

The strong positive correlation between national testing, wastewater monitoring data and Google RSVs has previously been demonstrated in the USA [34]. The relation of search term data to SARS-CoV-2 prevalence in wastewater changed over time, suggesting that such approaches require monitoring and constant evaluation, again suggesting that an approach combining data types may be optimal [34]. Importantly, the predictions made based on qPCR-based wastewater monitoring were marginally insignificantly related to recorded cases. Given the relative objectivity of this molecular monitoring, this is likely to reflect the inconsistent accuracy of national case data reporting as the pandemic progressed, highlighting the need for objective measures of virus prevalence irrespective of public participation. While these different data types dynamically interact and often imperfectly reflect one another, as demonstrated by our univariate predictions, together they could generate models with greater predictive power for forecasting improved above that of univariate approaches [34]. This aligns with the ‘One Health’ perspective of integrating different data types across disciplinary boundaries to monitor healthcare and epidemiological events more holistically [22,23]. Wastewater monitoring has been integrated into One Health frameworks for pathogen monitoring [52] and emerging concepts like antimicrobial resistance in the environment [53]. Given that infoveillance similarly aligns with the principles of One Health [23], this presents an ideal opportunity to integrate different data types for socio-biological monitoring of SARS-CoV-2 and other pandemic agents.

Limitations

Regarding infoveillance, this study relied exclusively on Google search volume data; while this represents the most used search engine and thus the greatest single source of such data, other search engines are regularly used that might provide different insights. Online search data, while an asset for assessing public responses, is also collected without the context of its users' motives, thus assumptions cannot reliably be made about the specific interests related to each search string. Even without this context, however, the search volumes presented in this study indicate interest, positive or negative, in those topics. Previous studies have demonstrated that the efficacy of these data in predicting epidemiological trends can be, at best, variable and, at worst, ineffective [13–15]; this can be mitigated to some degree via robust statistical methods to increase the reliability and accuracy of infoveillance for epidemiological 'nowcasting' [15] but integration of these data into more holistic frameworks across disciplinary boundaries could further ameliorate these inaccuracies and provide increasingly accurate predictions [22,23].

Whilst the qPCR data in this study represent a nationwide effort to monitor SARS-CoV-2, they do not comprehensively cover the nation of Wales which is otherwise fully represented by the Google Trends and national reporting data. Importantly, the qPCR data do account for all of South Wales which, in turn, accounts for approximately 71 % of the national population [54], meaning that these data should accurately reflect the overall national SARS-CoV-2 prevalence. Future studies could investigate how different spatiotemporal resolutions of data affect the accuracy and outcomes of analyses such as these, especially given that this will impact feasibility of long-term monitoring using most methods.

The progression of COVID-19 as a global pandemic continues to be extremely complicated and unpredictable, and the findings of this study focus on just one period in this evolving situation, prior to the emergence of the SARS-CoV-2 Omicron variant and its sub-lineages. More importantly, the early months of the pandemic are not represented due to the unavailability of qPCR data for that period. While this study relates primarily to those later months of the first year of the pandemic through to the second year, the use of Google Trends data may have been more powerful in the early months of the pandemic when public familiarity was lower and more people were seeking information.

Conclusions

Both molecular monitoring of wastewater and infoveillance approaches demonstrate potential for monitoring and prediction of epidemiological trends. Personal testing and surveys can introduce latency to monitoring, lack randomisation and can receive reduced participation for fear of positive test outcomes [10], thus reduced dependency on these data through widespread adoption of wastewater monitoring will likely improve the accuracy of epidemiological data. Wastewater monitoring has previously correlated strongly with national case data [32], but any decrease in this correlation must importantly be viewed with respect to public interest and how this might impact reported case data. Disease surveillance via wastewater monitoring provides many potential benefits, not least its objectivity. As public interest in the pandemic wanes, widespread molecular analysis of wastewater will become increasingly important as personal testing data become increasingly inaccurate at the population level. Public access to wastewater monitoring data has been facilitated through online reporting, including the data used in this study [38], but accessible presentation of these data in interactive dashboards, as has been the case for other national data, may increase public understanding, appreciation and use of this important data source.

Authors contributions:

JPC, SD and SW conceptualised the study. SD carried out the qPCR testing of wastewater. JPC and RABG analysed the data. JPC prepared figures. PK, AJW and DLJ oversaw the study. All authors contributed to the writing and editing of the manuscript.

Declarations:

Competing interests (e.g., any conflicting roles with companies etc): None of the authors have any conflicts of interest to declare.

Funding:

This work was supported by a grant from Welsh Government under the Welsh Wastewater Programme (C035/2021/2022). We also thank Tony Harrington at Dŵr Cymru-Welsh Water alongside staff at the wastewater treatment facility for their support in collecting the samples.

References

1. McLellan A, Godlee F. Covid 19: Christmas relaxation will overwhelm services. BMJ 2020;371:m4847.
2. Lange KW. The prevention of COVID-19 and the need for reliable data. Journal of Disease Prevention and Health Promotion 2020;4:39–52.
3. McDonald SA, Van Den Wijngaard CC, Wielders CCH, Friesema IHM, Soetens L, Paolotti D, Van Den Hof S, Van Hoek AJ. Risk factors associated with the incidence of self-reported COVID-19-like illness: data from a web-based syndromic surveillance system in the Netherlands. Epidemiol Infect 2021;149:e129. doi: 10.1017/S0950268821001187
4. McNeil C, Verlander S, Divi N, Smolinski M. The Landscape of Participatory Surveillance Systems Across the One Health Spectrum: Systematic Review. JMIR Public Health Surveill 2022 Aug 5;8(8):e38551. doi: 10.2196/38551
5. Leal-Neto O, Egger T, Schlegel M, Flury D, Sumer J, Albrich W, Babouee Flury B, Kuster S, Vernazza P, Kahlert C, Kohler P. Digital SARS-CoV-2 Detection Among Hospital Employees: Participatory Surveillance Study. JMIR Public Health Surveill 2021 Nov 22;7(11):e33576. doi: 10.2196/33576

6. Leal-Neto OB, Santos FAS, Lee JY, Albuquerque JO, Souza WV. Prioritizing COVID-19 tests based on participatory surveillance and spatial scanning. International Journal of Medical Informatics 2020 Nov;143:104263. doi: 10.1016/j.ijmedinf.2020.104263

7. Pandit JA, Radin JM, Quer G, Topol EJ. Smartphone apps in the COVID-19 pandemic. Nat Biotechnol 2022 Jul;40(7):1013–1022. doi: 10.1038/s41587-022-01350-x

8. Drew DA, Nguyen LH, Steves CJ, Menni C, Freydin M, Varsavsky T, Sudre CH, Cardoso MJ, Ourselin S, Wolf J, Spector TD, Chan AT, COPE Consortium, Chan AT, Drew DA, Nguyen LH, Joshi AD, Guo C-G, Ma W, Lo C-H, Mehta RS, Kwon S, Sikavi DR, Magicheva-Gupta MV, Fatehi ZS, Flynn JJ, Leonardo BM, Albert CM, Andreotti G, Beane-Freeman LE, Balasubramanian BA, Brownstein JS, Bruinsma F, Cowan AN, Deka A, Ernst ME, Figueiredo JC, Franks PW, Gardner CD, Ghobrial IM, Haiman CA, Hall JE, Deming-Halverson SL, Kirpach B, Lacey JV, Le Marchand L, Marinac CR, Martinez ME, Milne RL, Murray AM, Nash D, Palmer JR, Patel AV, Rosenberg L, Sandler DP, Sharma SV, Schurman SH, Wilkens LR, Chavarro JE, Eliassen AH, Hart JE, Kang JH, Koenen KC, Kubzansky LD, Mucci LA, Ourselin S, Rich-Edwards JW, Song M, Stampfer MJ, Steves CJ, Willett WC, Wolf J, Spector T. Rapid implementation of mobile technology for real-time epidemiology of COVID-19. Science 2020 Jun 19;368(6497):1362–1367. doi: 10.1126/science.abc0473

9. Menni C, Valdes AM, Freidin MB, Sudre CH, Nguyen LH, Drew DA, Ganesh S, Varsavsky T, Cardoso MJ, El-Sayed Moustafa JS, Visconti A, Hysi P, Bowyer RCE, Mangino M, Falchi M, Wolf J, Ourselin S, Chan AT, Steves CJ, Spector TD. Real-time tracking of self-reported symptoms to predict potential COVID-19. Nat Med 2020 Jul;26(7):1037–1040. doi: 10.1038/s41591-020-0916-2

10. Murakami M, Hata A, Honda R, Watanabe T. Letter to the editor: wastewater-based epidemiology can overcome representativeness and stigma issues related to COVID-19. Environmental Science & Technology 2020;54(9):5311.

11. Badell-Grau RA, Cuff JP, Kelly BP, Waller-Evans H, Lloyd-Evans E. Investigating the prevalence of reactive online searching in the COVID-19 pandemic: infoveillance study. Journal of Medical Internet Research 2020;22(10):e19791.

12. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009 Feb;457(7232):1012–1014. doi: 10.1038/nature07634

13. Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, Igusa T, Rothman RE. Influenza Forecasting with Google Flu Trends. Viboud C, editor. PLoS ONE 2013 Feb 14;8(2):e56176. doi: 10.1371/journal.pone.0056176

14. Lazer D, Kennedy R. What We Can Learn From the Epic Failure of Google Flu Trends. Wired 2015 Oct 1; Available from: https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/#:~:text=And%20then%2C%20GFT%20failed%E2%80%94and,the%20foibles%20of%20big%20data

15. Kandula S, Shaman J. Reappraising the utility of Google Flu Trends. Segata N, editor. PLoS Comput Biol 2019 Aug 2;15(8):e1007258. doi: 10.1371/journal.pcbi.1007258

16. Rovetta A, Bhagavathula AS. COVID-19-related web search behaviors and infodemic attitudes in Italy: Infodemiological study. Journal of Medical Internet Research 2020;22(5):1–10. doi: 10.2196/19374

17. Walker A, Hopkins C, Surda P. The use of Google Trends to investigate the loss of smell related searches during COVID-19 outbreak. International Forum of Allergy & Rhinology 2020;10:839–847. PMID:32279437

18. Ortiz-Martínez Y. Can Google trends predict COVID-19 incidence and help preparedness? The situation in Colombia. Travel Medicine and Infectious Disease 2020;37:101703.

19. Saegner T, Austys D. Forecasting and surveillance of COVID-19 spread using Google Trends: literature review. International Journal of Environmental Research and Public Health 2022;19(9):12394.

20. Effenberger M, Kronbichler A, Shin J Il, Mayer G, Tilg H, Perco P. Association of the COVID-19 pandemic with internet search volumes: a Google Trends analysis. International Journal of Infectious Diseases 2020;95:192–197. PMID:32305520

21. Zinsstag J, Schelling E, Waltner-Toews D, Tanner M. From “one medicine” to “one health” and systemic approaches to health and well-being. Preventive Veterinary Medicine 2011 Sep;101(3–4):148–156. doi: 10.1016/j.prevetmed.2010.07.003

22. Mackenzie JS, Jeggo M. The One Health Approach—Why Is It So Important? TropicalMed 2019 May 31;4(2):88. doi: 10.3390/tropicalmed4020088

23. Benis A, Tamburis O, Chronaki C, Moen A. One Digital Health: A Unified Framework for Future Health Ecosystems. J Med Internet Res 2021 Feb 5;23(2):e22189. doi: 10.2196/22189

24. Gundy PM, Gerba CP, Pepper IL. Survival of coronaviruses in water and wastewater. Food and Environmental Virology 2009;1:10.

25. Wade MJ, Jacomo A Lo, Armenise E, Brown MR, Bunce JT, Cameron GJ, Fang Z, Farkas K, Gilpin DF, Graham DW, Grimsley JMS, Hart A, Hoffmann T, Jackson KJ, Jones DL, Lilley CJ, McGrath JW, McKinley JM, McSparron C, Nejad BF, Morvan M, Quintela-Baluja M, Roberts AMI, Singer AC, Souque C, Speight VL, Sweetapple C, Walker D, Watts G, Weightman A, Kasprzyk-Hordern B. Understanding and managing uncertainty and variability for wastewater monitoring beyond the pandemic: Lessons learned from the United Kingdom national COVID-19 surveillance programmes. Journal of Hazardous Materials 2022;424:127456.

26. Ni G, Lu J, Maulani N, Tian W, Yang L, Harliwong I, Wang Z, Mueller J, Yang B, Yuan Z, Hu S, Guo J. Novel multiplexed amplicon-based sequencing to quantify SARS-CoV-2 RNA from wastewater. Environmental Science & Technology Letters 2021;8(8):683–690.

27. Bogler A, Packman A, Furman A, Gross A, Kushmaro A, Ronen A, Dagot C, Hill C, Vaizel-Ohayon D, Morgenroth E, Bertuzzo E, Wells G, Kiperwas HR, Horn H, Negev I, Zucker I, Bar-Or I, Moran-Gilad J, Balcazar JL, Bibby K, Elimelech M, Weisbrod N, Nir O, Sued O, Gillor O, Alvarez PJ, Crameri S, Arnon S, Walker S, Yaron S, Nguyen TH, Berchenko Y, Hu Y, Ronen Z, Bar-Zeev E. Rethinking wastewater risks and monitoring in light of the COVID-19 pandemic. Nature Sustainability 2020;3:981–990.

28. Hillary LS, Farkas K, Maher KH, Lucaci A, Thorpe J, Distaso MA, Gaze WH, Paterson S, Burke T, Connor TR, McDonald JE, Malham SK, Jones DL. Monitoring SARS-CoV-2 in municipal wastewater to evaluate the success of lockdown measures for controlling COVID-19 in the UK. Water Research 2021;200:117214.

29. Xu Y, Li X, Zhu B, Liang H, Fang C, Gong Y, Guo Q, Sun X, Zhao D, Shen J, Zhang H, Liu H, Xia H, Tang J, Zhang K, Gong S. Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding. Nature Medicine 2020;26:502–505.

30. Usman M, Farooq M, Hanna K. Existence of SARS-CoV-2 in wastewater: implications for its environmental transmission in developing communities. Environmental Science & Technology 2020;54(13):7758–7759.

31. He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, Lau YC, Wong JY, Guan Y, Tan X, Mo X, Chen Y, Liao B, Chen W, Hu F, Zhang Q, Zhong M, Wu Y, Zhao L, Zhang F, Cowling BJ, Li F, Leung GM. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nature Medicine 2020;26:672–675.

32. Medema G, Heijnen L, Elsinga G, Italiaander R, Brouwer A. Presence of SARS-Coronavirus-2 RNA in sewage and correlation with reported COVID-19 prevalence in the early stage of the epidemic in The Netherlands. Environmental Science & Technology Letters 2020;7(7):511–516.

33. Naughton CC, Roman FAJ, Alvarado AGF, Tariqi AQ, Deeming MA, Bibby K, Bivins A, Rose JB, Medema G, Ahmed W, Katsivelis P, Allan V, Sinclair R, Zhang Y, Kinyua MN. Show us the data: global COVID-19 wastewater monitoring efforts, equity, and gaps. medRxiv 2021;2021.03.14.21253564.

34. Liu Z, Jiang Z, Kip G, Snigdha K, Xu J, Wu X, Khan N, Schultz T. An infodemiological framework for tracking the spread of SARS-CoV-2 using integrated public data. Pattern Recognition Letters 2022;158:133–140.

35. Farkas K, Hillary LS, Thorpe J, Walker DI, Lowther JA, McDonald JE, Malham SK, Jones DL. Concentration and quantification of SARS-CoV-2 RNA in wastewater using polyethylene glycol-based concentration and qRT-PCR. Methods and Protocols 2021;4(1):17.

36. Oberacker P, Stepper P, Bond DM, Höhn S, Focken J, Meyer V, Schelle L, Sugrue VJ, Jeunen G-J, Moser T, Hore SR, Meyenn F von, Hipp K, Hore TA, Jurkowski TP. Bio-On-Magnetic-Beads (BOMB): Open platform for high-throughput nucleic acid extraction and manipulation. PLoS Biology 2019;17(1):e3000107.

37. Vogels CBF, Brito AF, Wyllie AL, Fauver JR, Ott IM, Kalinich CC, Petrone ME, Casanovas-Massana A, Muenker MC, Moore AJ, Klein J, Lu P, Lu-Culligan A, Jiang X, Kim DJ, Kudo E, Mao T, Moriyama M, Oh JE, Park A, Silva J, Song E, Takahashi T, Taura M, Tokuyama M, Venkataraman A, Weizman O-E, Wong P, Yang Y, Cheemarla NR, White EB, Lapidus S, Earnest R, Geng B, Vijayakumar P, Odio C, Fournier J, Bermejo S, Farhadian S, Cruz CS Dela, Iwasaki A, Ko AI, Landry ML, Foxman EF, Grubaugh ND. Analytical sensitivity and efficiency comparisons of SARS-CoV-2 RT–qPCR primer–probe sets. Nature Microbiology 2020;5:1299–1305.

38. Stachler E, Kelty C, Sivaganesan M, Li X, Bibby K, Shanks O. Quantitative CrAssphage PCR assays for human fecal pollution measurement. Environmental Science & Technology 2017;51(16):9146–9154.

39. Kitajima M, Tohya Y, Matsubara K, Haramoto E, Utagawa E, Katayama H. Chlorine inactivation of human norovirus, murine norovirus and poliovirus in drinking water. Letters in Applied Microbiology 2010;51:119–121.

40. Wilde H, Perry WB, Jones O, Kille P, Weightman A, Jones DL, Cross G, Durance I. Accounting for dilution of SARS-CoV-2 in wastewater samples using physico-chemical markers. Water 2022;14:2885.

41. UK Government. Gov.UK Coronavirus (COVID-19) in the UK. GovUK. 2021. p. 1. Available from: https://coronavirus.data.gov.uk/details/cases [accessed Dec 1, 2021]

42. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020.

43. Cuff JP, Dighe SN, Watson SE, Badell-Grau RA, Weightman AJ, Jones DL, Kille P. An infoveillance analysis of public interest, national data and wastewater monitoring in Wales, UK. Zenodo; 2022. doi: 10.5281/ZENODO.7260269

44. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen T, Miller E, Bache S, Müller K, Ooms J, Robinson D, Seidel D, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H. Welcome to the Tidyverse. JOSS 2019 Nov 21;4(43):1686. doi: 10.21105/joss.01686

45. Harrell F. Hmisc: Harrell Miscellaneous. 2020.

46. Wei T. R package “corrplot”: visualization of a correlation matrix. 2017.

47. Garnier S. viridis: default color maps from “matplotlib.” 2018.

48. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag; 2016.

49. Neuwirth E. RColorBrewer: ColorBrewer palettes. 2014.

50. Husain I, Briggs B, Lefebvre C, Cline DM, Stopyra JP, O’Brien MC, Vaithi R, Gilmore S, Countryman C. Fluctuation of public interest in COVID-19 in the United States: retrospective analysis of Google Trends search data. Journal of Medical Internet Research Public Health and Surveillance 2020;6(3):e19969.

51. Nindrea RD, Sari NP, Lazuardi L, Aryandono T. Validation: the use of Google Trends as an alternative data source for COVID-19 surveillance in Indonesia. Asia Pacific Journal of Public Health 2020;32(6–7):368–369.

52. Xiao K, Zhang L. Wastewater pathogen surveillance based on One Health approach. The Lancet Microbe 2023 May;4(5):e297. doi: 10.1016/S2666-5247(23)00039-3

53. Miłobedzka A, Ferreira C, Vaz-Moreira I, Calderón-Franco D, Gorecki A, Purkrtova S, Jan Bartacek, Dziewit L, Singleton CM, Nielsen PH, Weissbrodt DG, Manaia CM. Monitoring antibiotic resistance genes in wastewater environments: The challenges of filling a gap in the

One-Health cycle. Journal of Hazardous Materials 2022 Feb;424:127407. doi: 10.1016/j.jhazmat.2021.127407

54. Llywodraeth Cymru Welsh Government. Summary statistics by regions of Wales: 2020. Llywodraeth Cymru Welsh Government. 2020. Available from: https://www.gov.wales/summary-statistics-regions-wales-2020