Standardizing data reporting in the research community to enhance the utility of open data for SARS-CoV-2 wastewater surveillance†


SARS-CoV-2 RNA detection in wastewater is being rapidly developed and adopted as a public health monitoring tool worldwide. With wastewater surveillance programs being implemented across many different scales and by many different stakeholders, it is critical that data collected and shared are accompanied by an appropriate minimal amount of meta-information to enable meaningful interpretation and use of this new information to enhance the utility of open data for wastewater surveillance.

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d1ew00235j

Water impact

Extensive wastewater surveillance data are being generated during the COVID-19 pandemic; however, there is no consensus on the meta-information that should be reported with wastewater SARS-CoV-2 concentrations. Complete and consistent data are important for regional, national, and international data synthesis. The minimum recommended meta-information here aims to set a framework for wastewater surveillance data reporting.
Information source and intercomparison across datasets. While some databases are being developed for specific surveillance programs locally, regionally, nationally, and internationally, common globally-adopted data standards have not yet been established within the research community. Establishing such standards will require national and international consensus on what meta-information should accompany SARS-CoV-2 wastewater measurements. To establish a recommendation on minimum information to accompany reporting of SARS-CoV-2 occurrence in wastewater for the research community, the United States National Science Foundation (NSF) Research Coordination Network on Wastewater Surveillance for SARS-CoV-2 hosted a workshop in February 2021 with participants from academia, government agencies, private companies, wastewater utilities, public health laboratories, and research institutes. This report presents the primary two outcomes of the workshop: (i) a recommendation on the set of minimum meta-information that is needed to confidently interpret wastewater SARS-CoV-2 data, and (ii) insights from workshop discussions on how to improve standardization of data reporting.

Introduction

Following early reports of SARS-CoV-2 RNA detection in sewage,1,2 there has been high interest in the application of wastewater surveillance for monitoring the COVID-19 pandemic. Many academic researchers, government agencies, and commercial scientists have developed methods for detecting SARS-CoV-2 in wastewater and applied these methods to inform COVID-19 pandemic public health response.3,4 Ensuring prompt, appropriate access to complete and organized data following FAIR data principles (findable, accessible, interoperable, and reusable) is critical for scientific advancement,5 and the COVID-19 pandemic has highlighted the utility of publicly available datasets, such as transit data for assessing lockdown impacts and genome sequencing data for tracking viral transmission dynamics.6,7 Large-scale wastewater surveillance efforts, including those in the United States,8,9 the European Union,10,11 Canada,12 Australia,13 and Turkey,14 are developing data reporting structures for their own internal databases. However, it is not always clear how reporting structures for these databases were developed, and many researchers, utilities, and public health officials are generating additional wastewater surveillance data outside of these programs.15 As such, there is not yet a consensus on what meta-information should accompany these measurements to enable a careful and judicious data interpretation, nor a central and open repository for all wastewater surveillance data, though efforts are underway to develop these resources.11,16

Here we provide initial guidance on minimum appropriate meta-information related to infrastructure characteristics, collection and processing procedures, and quantification methods to accompany SARS-CoV-2 wastewater surveillance. We recognize that specific data applications may require additional information depending on the purpose of a research study or surveillance program; however, our objective is that the guidance developed here, using an open community-led format and with input from many ongoing SARS-CoV-2 wastewater surveillance efforts, will advance a more standardized and accessible reporting protocol. This will enable more robust comparisons across studies and create more reusable and interoperable long-term resources for future applications of wastewater surveillance.

Methods

The Research Coordination Network (RCN) on Wastewater Surveillance for SARS-CoV-2 is a U.S. National Science Foundation (NSF) funded initiative launched in August 2020 to advance research and education in the detection and longitudinal monitoring of SARS-CoV-2 and COVID-19 via wastewater analysis. In February 2021, this NSF RCN convened a workshop with the goal of reaching an agreement on the minimum information that must accompany wastewater SARS-CoV-2 measurements for the data to be broadly useful for wastewater-based epidemiology applications. Participants in the workshop were nominated in response to an open call and subsequently selected to represent the wide array of technical backgrounds and expertise that is relevant to wastewater surveillance. The 28 participants represented various professional sectors, including academia, government agencies, private companies, wastewater utilities, public health laboratories, and research institutes, and included international perspective from four countries.

Prior to the workshop, moderators developed a list of 47 possible meta-information variables based on the U.S. Centers for Disease Control & Prevention (CDC) National Wastewater Surveillance System (NWSS)8 data reporting structure (Table S1†). As there are already well-established community guidelines on necessary data reporting for quantitative PCR (qPCR) and digital PCR (dPCR) – the MIQE17 and dMIQE18 guidelines, respectively – variables covered in these guidelines were specifically excluded from the workshop discussion. Using a survey, workshop participants were asked to rank each variable on a 5-point scale from “unnecessary” (1) to “essential” (5) based on the question: “How important is this variable for appropriate interpretation of SARS-CoV-2 wastewater monitoring data?” Participants were also provided the opportunity to suggest additional reporting variables, which resulted in suggestions of 23 new variables (Table S1†). During the workshop, participants were provided with the aggregate rankings of each variable. They were then divided into four groups (wastewater treatment plant & infrastructure, sample collection, sample processing, and target quantification), where groups discussed the preliminary rankings and identified a final set of variables within their category that are essential for interpreting SARS-CoV-2 wastewater surveillance data. Participants were asked to focus on only the minimum meta-information they would require to interpret an unfamiliar dataset and to consider practicality in...
measuring or obtaining the data for determining essential variables. Each group then presented the results of their discussion to the full set of workshop participants, explained their rationale, and incorporated contributions from other participants. This resulted in an initial agreement from the entire group of workshop participants. Following the workshop, preliminary variable rankings, discussion group rationale, and notes from workshop discussions were combined by workshop moderators to devise a final set of recommended minimum meta-information, which are described below.

Minimum meta-information for data reporting

Based on variable rankings and discussion during the convened workshop, we recommend a minimum set of information that should be included with SARS-CoV-2 wastewater measurements (Table 1). Here we explain why certain variables were selected and provide recommendations for the level of detail that should be included for each variable.

Wastewater treatment plant & infrastructure

Many wastewater surveillance efforts are focused on sampling at wastewater treatment plants, either at the primary influent or primary sludge locations, because these sites provide community-level coverage, are easily sampled, and are generally well-mixed. However, there are also important applications for wastewater surveillance in sewer regions upstream of treatment plants or at specific facilities or buildings. Regardless of the scale of sampling, it is critical to define the approximate population served and whether the sampling location is a combined or separated sewer system, as stormwater flows in combined sewers can dilute target waste streams and affect data interpretation. Reporting of mean daily flow rates is particularly helpful for estimating the population contributing to a sample, viral loading rates, or infection prevalence in the population. However, measuring or estimating flow rates at upstream sewer sites is often not feasible, and treatment plant influent flow rates are not directly relevant to sampling of primary sludge. We therefore encourage reporting of flow rates when possible but recognize that these data are often not available. While specific GPS coordinates of sample sites would be valuable for cross-referencing with other databases, this level of spatial specificity may lead to privacy concerns in some cases. Therefore, we determined that location information should be reported at the county or municipality level to allow comparison to other public health data while maintaining a degree of sampling anonymity. Wastewater-based surveillance can also be applied to non-sewered waste streams, such as septic tanks, pit latrines, or drainage ditches in areas lacking piped sanitation infrastructure. While our discussion focused specifically on sewered systems, the general scope of variables identified in Table 1 are likely still applicable and could be adapted to data collected from non-sewered systems.

Sample collection

The type of samples collected and the manner in which they are collected are critical for understanding the quality of data that can be obtained from a wastewater sample. The impacts of grab sampling versus composite sampling on resulting data utility are not yet clear, but prior work suggests that daily fluctuations in wastewater flows and possibly defecation rates may impact results. Therefore, identifying the type of sample (grab, composite), the duration of compositing, and the sampling date and time are important for data interpretation, as well as comparison to other public health data sources. During workshop discussion, participants noted that sampling dates are not uniformly reported, particularly for composite samples that can span multiple dates. This ambiguity is especially problematic for comparing wastewater-based data to other public health data sets. It is therefore recommended that laboratories collecting composite samples at minimum report the start date and start time of the composite sampling program, as this information, along with the sampling duration, identifies the complete time period captured by the composite sample. Additionally, when reporting the sample matrix collected, it should be noted whether the wastewater was collected after any pre-treatment, such as chlorination or ferric chloride addition, which may impact the results obtained by laboratory analysis of SARS-CoV-2 RNA.

Sample processing

If samples are not processed immediately, SARS-CoV-2 RNA targets may undergo decay during storage, and storage temperature can impact the extent of decay, especially if samples undergo freeze–thaw cycles. Laboratories should therefore report the temperature of sample storage prior to processing, including any freeze–thaw cycles. Exact temperatures during sample shipping may not be available, but qualitative shipping conditions (e.g., on ice, dry ice, refrigerated, etc.) should still be specified as available. Many different methods can be used to concentrate and isolate viral RNA from wastewater, and it is therefore important to identify the major processing approach and the results of negative processing controls (i.e., extraction blanks). We advocate for the inclusion of a reference to specific concentration and extraction protocols when reporting SARS-CoV-2 wastewater surveillance data, as protocol differences will be important for comparisons across laboratories. In cases where protocols are not yet published in the peer-reviewed literature or defined by a kit manufacturer, open-source resources such as protocols.io can be used to document and reference laboratory-specific protocols. While the utility of spiked-in recovery standards is not yet universally agreed upon, and SARS-CoV-2 concentration data can be interpreted without this information, we nonetheless recommend the additional reporting of recovery controls and recovery efficiency when possible to facilitate comparisons between studies, samples, and methods.
Finally, the amount of sample (volume or mass) analyzed has to be specified to enable the calculation of analyte concentrations in sewage and determination of methodological detection limits.

**Target quantification**

Most current wastewater surveillance efforts rely on reverse transcription qPCR (RT-qPCR) or reverse transcription dPCR (RT-dPCR) to quantify specific SARS-CoV-2 gene targets. As qPCR and dPCR experiments use different strategies for target quantification, it is important to identify the type of PCR used to measure SARS-CoV-2 concentrations and to report the minimal information that has been previously established for these quantification methods (i.e., MIQE17 and dMIQE18 standards). These standards include specifications for reporting on no-template controls, positive controls, assay efficiencies, limits of detection and quantification, and inhibition testing, which are all critical for appropriate interpretation of qPCR and dPCR results. For SARS-CoV-2 concentrations, we recommend reporting concentrations in terms of copies per liter of sewage or per gram dry weight of sewage solids without normalization to recovery controls or endogenous wastewater controls. Endogenous wastewater controls are additional targets within a sample that are associated with typical human inputs to sewer systems and may serve as both an indication of fecal matter content and as a recovery control.29 It is not yet clear exactly how recovery or endogenous control data should be specified that raw concentrations for SARS-CoV-2, recovery controls or endogenous wastewater controls.

## Table 1  List of minimum information to accompany measurements of SARS-CoV-2 in wastewater

<table>
<thead>
<tr>
<th>Category</th>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wastewater treatment plant &amp; infrastructure</td>
<td>Sample location type</td>
<td>Primary influent, primary sludge, street line manhole, pump station, septic,</td>
</tr>
<tr>
<td></td>
<td>Population served</td>
<td>on-facility (university campus, correctional facility, etc.), other</td>
</tr>
<tr>
<td></td>
<td>Combined or separated system?</td>
<td>Combined, separated, mixed</td>
</tr>
<tr>
<td></td>
<td>Primary county/municipality served</td>
<td>County/municipality, state/province, country</td>
</tr>
<tr>
<td></td>
<td>Flow</td>
<td>Mean daily flow on day(s) of for sample collection. List “N/A” if this information is not available (e.g. for sewer or building samples)</td>
</tr>
<tr>
<td>Sample collection</td>
<td>Sample collection type</td>
<td>Grab, composite (flow-weighted or time-weighted, including composite duration), other</td>
</tr>
<tr>
<td></td>
<td>Sample matrix</td>
<td>Raw wastewater, pre-treated wastewater (including pre-treatment type), wastewater solids, other</td>
</tr>
<tr>
<td></td>
<td>Sample date</td>
<td>Date (MM/DD/YYYY) of sample collection from sewer system; if composite, composite sampling start date</td>
</tr>
<tr>
<td></td>
<td>Sample time</td>
<td>Time of sample collection from sewer system; if composite, composite sampling start time</td>
</tr>
<tr>
<td>Sample processing</td>
<td>Pre-concentration storage temperature</td>
<td>Degrees Celsius (if available), on ice, dry ice, refrigerated, frozen. Specify number of freeze-thaw cycles, if any</td>
</tr>
<tr>
<td></td>
<td>Concentration method &amp; citation</td>
<td>PEG precipitation, ultrafiltration, none, HA filtration, ultracentrifugation, nanotrap beads, other; include protocol citation</td>
</tr>
<tr>
<td></td>
<td>Recovery control name &amp; efficiency</td>
<td>BCoV, BRSV, MHV, OC43, other; include recovery efficiency if a control was used. List “none” if no recovery control was used</td>
</tr>
<tr>
<td></td>
<td>Extraction method &amp; citation</td>
<td>Kit-based (include kit name), TRizol, MagBead, other; include protocol citation</td>
</tr>
<tr>
<td></td>
<td>Amount of sample processed</td>
<td>Starting volume [mL] or mass [g] of raw sample processed</td>
</tr>
<tr>
<td></td>
<td>Extraction blanks results</td>
<td>Signal not detected, signal detected (% positive), blanks not used qPCR, dPCR, other</td>
</tr>
<tr>
<td>Target quantification</td>
<td>PCR type</td>
<td>Concentration back-calculated to raw sample volume/mass basis (e.g. copies per L wastewater or copies per g dry weight sludge)</td>
</tr>
<tr>
<td></td>
<td>SARS-CoV-2 concentrations</td>
<td>Flag as below LOD (BLOD) or below LOQ (BLOQ)</td>
</tr>
<tr>
<td></td>
<td>Identification of samples below LOD or LOQ</td>
<td>Genie target and primers/assay</td>
</tr>
<tr>
<td></td>
<td>SARS-CoV-2 target gene(s)</td>
<td>PMMoV, crAssphage, HF183, other; include concentration if measured. List “none” if no endogenous wastewater control was used</td>
</tr>
<tr>
<td></td>
<td>Endogenous wastewater control name &amp; concentration</td>
<td>Includes specifications for assessing RNA quantification &amp; integrity, reaction conditions, no-template controls, positive controls, assay efficiencies (for qPCR), LOD, LOQ, inhibition testing, and others. See references17,18 for complete lists</td>
</tr>
</tbody>
</table>
|                                               | Required MIQE / dMIQE standards               | * Polyethylene glycol.  

*b* Bovine coronavirus.  
*i* Bovine respiratory syncytial virus.  
*^d^* Murine hepatitis virus.  
*^e^* Human coronavirus OC43.  
*^f^* Limit of detection.  
*^g^* Limit of quantification.  
*^h^* Pepper mild mottle virus.  
*^i^* Human Bacteroides marker HF183.
changes in analysis strategies. Finally, as is true for any dilute target, methodological limits of detection and quantification can substantially impact data interpretation, and identification of sample measurements that fall below these limits is therefore necessary.

Summary and perspective

The variables described in the previous section and summarized in Table 1 represent a recommendation of the minimum information that should be reported with SARS-CoV-2 wastewater concentration data. Table 1 also generally describes additional variables required by MIQE\textsuperscript{17} or dMIQE\textsuperscript{18} guidelines, and these guidelines must also be referenced to ensure complete data reporting. Additionally, as the research and knowledge continue to advance, future refinements of the recommendations and inclusion of additional meta-information may be warranted. For example, two types of processing controls — recovery controls and endogenous wastewater controls — are used by many laboratories to monitor method performance and may be useful for comparison across laboratories. We have included these controls in our recommended minimum meta-information to encourage their use and reporting, but also recognize that utility of these controls remains uncertain and data reported without these parameters is still useful for local surveillance with consistent methods. While our discussions focused on PCR methods, the overall structure of variables we have included in this guidance could easily be broadened or adapted to include other types of data.

The final set of minimum required variables in Table 1 is similar to the required variables for reporting in the CDC NWSS database. While the initial set of variables provided during the workshop was based on the CDC NWSS data reporting structure, workshop participants were not provided with any additional information on NWSS data requirements. By arriving at a similar set of variables, our recommendation reinforces existing data standards for wastewater surveillance and provides a useful framework for laboratories to share their data in a way that will improve interoperability across datasets and databases.

Conclusion

We recommend that laboratories include the minimum data listed in Table 1 when reporting SARS-CoV-2 RNA measurements in wastewater, whether for scientific publication or public dashboards. We also encourage laboratories to make their data publicly available whenever feasible, ideally through deposition into public repositories, as this can greatly facilitate efficient technology development and method optimization. Some wastewater surveillance data may be subject to non-disclosure agreements or other sharing restrictions based on privacy concerns, and laboratories should work in collaboration with institutional review boards, health agencies, and other stakeholders to carefully ensure ethical data sharing. The framework provided here is purposefully simple and can be modified to accommodate different circumstances. We also stress that this framework is meant to address the minimum meta-information necessary for reporting only. Additional meta-information is valuable and may indeed be necessary for more complex data applications, such as modeling. Wastewater surveillance is a rapidly developing technique with applications beyond the present COVID-19 pandemic. As academic researchers, government agencies, and private companies continue to innovate and invest in this technology, the framework provided here can serve as a basis for harmonizing data reporting across applications.

Disclaimer

The research presented was not performed or funded by EPA and was not subject to EPA’s quality system requirements. The views expressed in this article are those of the author(s) and do not necessarily represent the views or the policies of the U.S. Environmental Protection Agency.

Author contributions


Conflicts of interest

C. D. is an employee of Biobot Analytics, Inc. P. K. is the founder of Ventic Technologies. B. M. S. is an employee of IDEXX Laboratories, Inc. R. U. H. is a cofounder of AquaVitas, LLC and the nonprofit project OneWaterOneHealth of the Arizona State University Foundation.

Acknowledgements

This work was supported by National Science Foundation grants 2038087 (J. S. M., A. B. B., R. U. H., K. B., J. D. V.), 2027752 (K. B.), 2023057 (A. B. B.), 2029025 (J. D. V.), and 2028564 (R. U. H.). We would like to thank Aaron Bivins for contributing to initial variable rankings and Angela Coulliette-Salmond and Wiley Jennings for their contributions to variable rankings and workshop discussions.

References


