Integrating Wastewater and Public Health Data

AUTHORS
Lauren Anderson*, Heather Ness, Rochelle Holm, Rebecca Schneider, and Ted Smith

* Corresponding author information:
Christina Lee Brown Envirome Institute
University of Louisville
Lauren.anderson@louisville.edu

By bringing together data from multiple sources in a structured way, public health officials can fill gaps when one data source is weak or see nuanced insights that might otherwise go unnoticed. Data integration allows for analysis that makes it easier to act on data. For instance, plots that show trends across wastewater and clinical data and risk scores that show where to prioritize resources can serve as justification for decisions. To be successful, integration for wastewater and public health data should be timely, consistent, disseminated broadly, and enable public officials to make more informed decisions.

Integrate data to fill gaps

Individually, wastewater data and public health data provide insights into only a portion of the community-level burden of COVID-19, seasonal flu, gastrointestinal bugs, and other diseases. For example, with individual case count data, infections are only registered if people first have symptoms and then visit a doctor or testing site. If either of these steps does not occur, we miss important data that helps us understand the full scope and context of illness. Although wastewater cannot provide local health agencies with the type of individual-level data contained in administratively reported case counts from hospitals, it can help officials understand viral infection levels and spread across communities without relying on individual testing practices. When paired with complementary data (see Box 1), wastewater data provide a more comprehensive view of disease burden in a community and can be more useful as a public health surveillance tool.

Box 1. Complementary data sources

Wastewater data (from sanitation agency or wastewater laboratory): target pathogen levels (SARS-CoV-N1, Influenza, etc.), human biomarkers (PMMoV, CRASSPHG, etc.), toxins, volatile organic chemicals, GIS sewer network attributes, effluent flow, service population

Public health data from public health departments or hospital systems): test positivity rate, COVID-19 incidence, hospitalizations, intensive care unit bed occupancy, mortality, vaccination rates

Other data (from census, ACS, or ArcGIS databases): age, income, race, urban or rural classification, underlying health conditions, workforce participation, community features (apartments versus standalone housing, presence of prisons, universities, nursing homes, or other communal living facilities), land use, population mobility, community events, and tourism data

Assess and access available data

The first step in integrating wastewater data with other public health data is to assess what local data are already available (see Box 1 for examples) and understand the processes that enable sharing the data.

1/ **Characterize existing data.** Public health agencies might have access to a variety of health metrics, such as positive cases, hospitalization rates, intensive care unit bed occupancy, vaccinations, and deaths. It is important to know how often each data set is collected and at what scale (for example, by address, zip code, or county) to assess the potential added value of wastewater data.

Tips to keep in mind when acquiring and assessing data:

/ Weekends and holidays can cause lags in data reporting that result in artificial variability of certain measures, such as case rates and hospitalizations. Smoothing the data (for example, by calculating rolling seven-day averages) can help you better interpret data.

/ Timing (date) and scale (location) are the two basic shared characteristics that allow you to merge distinct data sets.

/ Knowing where data originate and what entities have ownership or interest in the data (for example, hospitals, states, or other elected officials) is important.

/ Special attention must be given when data includes personal health identifiers such as names, addresses, dates of treatment, birth dates, and so on. Even if no single data set includes personal health identifiers, take care when overlaying or combining multiple data sets together to ensure that no one can triangulate identifying information from the component data sources.

/ If public health data must be aggregated to neighborhoods, GIS shape files need to be shared with the health agency during initial partnership discussions.

Mathematica® Progress Together
The Rockefeller Foundation

2/ **Determine what insights you hope to gain.** For instance, you might ask whether there is a relationship between viral concentrations in wastewater and clinical measures, such as the number of positive tests, positivity rate, or hospital admissions. By plotting these metrics together, you can uncover patterns. For example, if the number of hospitalizations spike two weeks after wastewater viral concentration levels spike, then officials can make health care staffing decisions informed by this relationship (that is, the potential early warning from the wastewater data). Meeting with biostatisticians and epidemiologists at the outset will be helpful because they can share limitations for data use or suggest ways to handle missing data.

3/ **Access the data**. If the data are already publicly available (for example, through a dashboard), you must understand any restrictions on reuse of the data. If the data are not publicly available, data sharing agreements must be established before sharing any data. These agreements can take several weeks or up to several months to be fully executed. Memorandums of understanding or data transfer agreements outline what data will be received, processes for receiving the data, data storage, and how the data will be used, including later publication. Keep in mind that there are many levels of stakeholders in any given institution, and you should create a stakeholder list in the initial partnership meetings. For example, personnel from legal and risk departments often review data agreements before they are executed. And if the wastewater activity is being undertaken in partnership with a research study— particularly if personal health identifiers are being shared— the completed data sharing agreement must be submitted to and accepted by the university's institutional review board. One key to success and speedy execution is to have buy-in from a high-level champion, such as a medical director at the public health agency, who can help get the stakeholders to agree to data sharing. Finally, data managers from the organizations sharing data should know about the data sharing agreement and what is expected of them. After all agreements are signed, the data transfer can begin by making a formal request to the public health data manager.

Case studies from Louisville, Kentucky

In Louisville, the fact that both wastewater and clinical testing results rose shortly after the 2021 holiday season provided decision makers with confidence in the wastewater data. Examining wastewater data alongside clinical testing data also highlighted the increased virility of the novel Omicron variant compared with the Delta variant. Further, by pairing wastewater and clinical testing data, researchers could identify neighborhoods where under-testing was likely occurring by looking for areas where wastewater viral concentrations were elevated but case counts were not. Louisville's public health agency used such information to deploy targeted testing resources to those areas.

Prepare the data for integration

Before integrating data, check the quality of the received public health data and normalize wastewater results based on human biomarkers. Then, integrate the data based on shared basic characteristics such as date or location.

4/ **Confirm data quality.** Ensure that the received data are reliable (that is, complete, accurate, and timely), and track any gaps, potential data inaccuracies, limits of detection, and suspicious outliers (see Box 2). Ensure that column headers and cell formatting across spreadsheets match before beginning integration. For example, transform all dates into the short date format (for example, 12/12/2022) from long or free formats. Ensure that there are no spaces in column headings or spreadsheet titles. The public health data set should come with a data dictionary that explains heading titles and notes about the data. Layering data might require integration across software, such as from spreadsheets to GIS files.

Box 2. Data quality

According to the Centers for Disease Control and Prevention, there are three key elements of data quality for public health surveillance: completeness, accuracy, and timeliness. Data are considered complete when they capture all the cases of interest. Data are accurate when the information they reflect is true. Data are timely when they are available within the period of time when it is useful. The Centers for Disease Control and Prevention offers simple tools to improve data quality.

5/ **Normalize wastewater data.** Normalization is the process of structuring data so that you can compare measures across sites or data sets. Normalizing wastewater data is important because the amount of viral material in samples fluctuates with the number of people contributing to the sewer system, wastewater flow rate, dilution from rainfall, and other environmental factors. Fecal indicators, such as Pepper Mild Mottle Virus (PMMoV), can serve to normalize or adjust the results based on the amount of human-made material in the sewer system. The University of Louisville normalizes SARS-CoV-2 concentrations in wastewater samples by dividing the number of SARS genetic copies (N1) by the number of PMMoV copies in the sample (using the formula N1/PMMoV). Read more about how fecal indicators can serve to calibrate epidemiological models for pathogen surveillance. See Box 3 for other normalization approaches.

Box 3. Normalization approaches

For wastewater data to be included on the Centers for Disease Control and Prevention's National Wastewater Surveillance System Dashboard, data must be normalized by flow rate and service population size. Flow rate and service population size normalization are easier because they can be calculated through demographic and spatial data. Not all sewer collection points have flow rate data. Population size normalization will not yield accurate results if the number of people contributing to the sewershed changes because of tourism, weekday commuters, and so on. Learn more about normalization for the Centers for Disease Control and Prevention's National Wastewater Surveillance System Dashboard.

Integrate wastewater and public health data

With two prepared data sets that share a common merge field (usually date or location), you can begin integration.

6/ Use Excel, ARCGIS Pro, or other software to merge the data sets together.

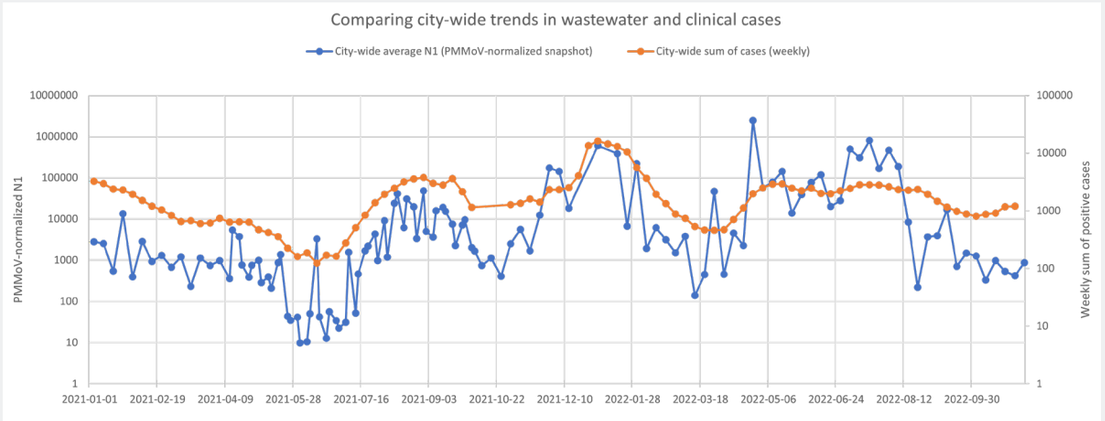
7/ **Create visualizations.** Now that you have merged the data, you should create easily understandable graphs to help public health leaders identify patterns and relationships in the wastewater and public health metrics. There are many ways to visualize the data, at varying scales, such as citywide (Figure 1), within individual neighborhoods (Figure 2), or by sewersheds served by water quality treatment centers. Each data point can reflect daily counts, moving averages, or weekly averages.

Share integrated data

One way to leverage integrated data is through a decision framework that assigns relative risk to areas based on several metrics, wastewater level, wastewater trends, case rate, case trends, and vaccine rate. This epidemiological and geographically contextual model is the vehicle for translating integrated wastewater and public health data to city officials.

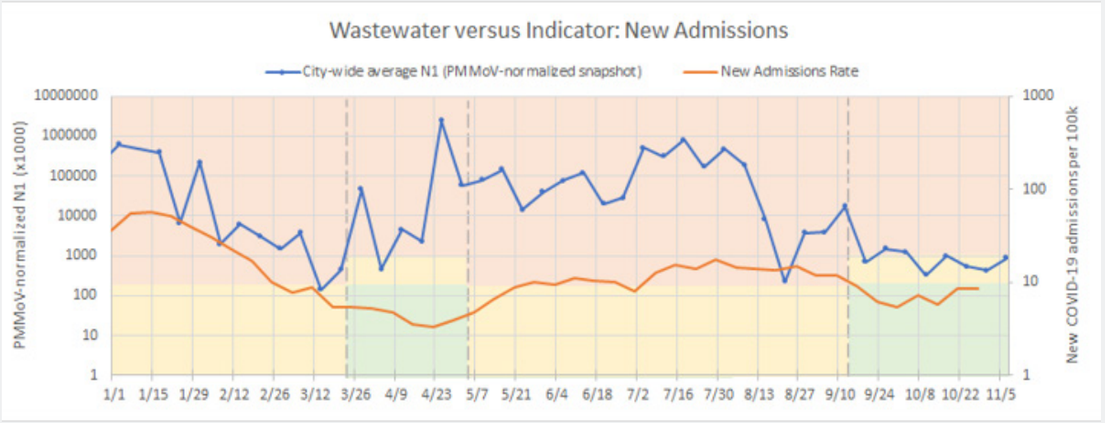
8/ **Develop an integrated risk score.** Decide collaboratively with your public health agency what key metrics are important to track. In Louisville, case rate was the most important metric for decision-making at the city level. Because of the wave pattern of COVID-19 infection, we included increasing and decreasing trends for wastewater and positive cases. After you decide on metrics, you can assign scores based on data and rank areas by relative risk of infection.

Figure 1. Plot of wastewater viral concentrations overlaid with clinical case counts across Louisville



Here, we plot wastewater and positive clinical cases across Louisville-Jefferson County together. When the wastewater (N1) line is much higher than the clinical case line, this might warrant further investigation. For example, when we see this pattern, we might ask whether wastewater is a more reliable measure of community infection than clinical testing alone.

Figure 2. New hospitalizations plotted alongside wastewater results for the entire city of Louisville



Notes: Color coding is based on the Centers for Disease Control and Prevention's Indicators for Monitoring COVID-19 Community Levels. Here, we plot wastewater and hospital metrics together. When we see a wastewater spike, such as the period from 4/1/2022 to 5/1/2022, and then a subsequent spike in new admissions around three weeks later, public health officials might advise health systems to increase staffing in the weeks after another wastewater spike based on this past pattern.

Figure 3 shows a visualization of integrated data and the scores for each of the key metrics for the Cedar Creek neighborhood. Trends are apparent in the three graphs at the bottom. Figure 4 shows the standardization (see Box 4) of all the key metrics and the weighting given to case rate. By using Figures 3 and 4, we can calculate the risk score for Cedar Creek:

Wastewater level: 3/5 = 0.60
Wastewater trend: 2/3 = 0.67
Case rate: 5/8 = 0.62 x 2 = 1.25
Case trend: 1/3 = 0.33
Vaccine coverage: 3/3 = 1
Sum for risk score: 3.85

Box 4. Standardization

When combining multiple metrics, such as wastewater level, wastewater trend, case rate, case trend, and vaccine coverage into a single score, it is important to standardize each metric. In this case, we put each metric on a 0 to 1 scale and then assigned double weight to the metric that was most important to Louisville's health department: case rate. By standardizing each of the five metrics to 1, the resulting risk scores will range from 0-6, allowing for ranking based on risk.

Conclusion

Combining information from multiple sources provides decision makers with a more holistic picture of disease dynamics and greater confidence to act. The discrete data sets can act as validators for each other or alert decision makers to changes during evolving public health situations. Wastewater monitoring should be viewed as another public health tool to understand population health, especially in light of its ability to illuminate patterns of infection when integrated with demographic and geographic data.

Suggested citation: Anderson, L., H. Ness, R. Holm, R. Schneider, and T. Smith. "Integrating Wastewater and Public Health Data" Washington, DC: Mathematica, 2022.

Acknowledgements and funding: This brief is based on research funded by The Rockefeller Foundation and was prepared by and with Mathematica. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of The Rockefeller Foundation.

Figure 3. Sample risk-decision framework summary from the Cedar Creek neighborhood

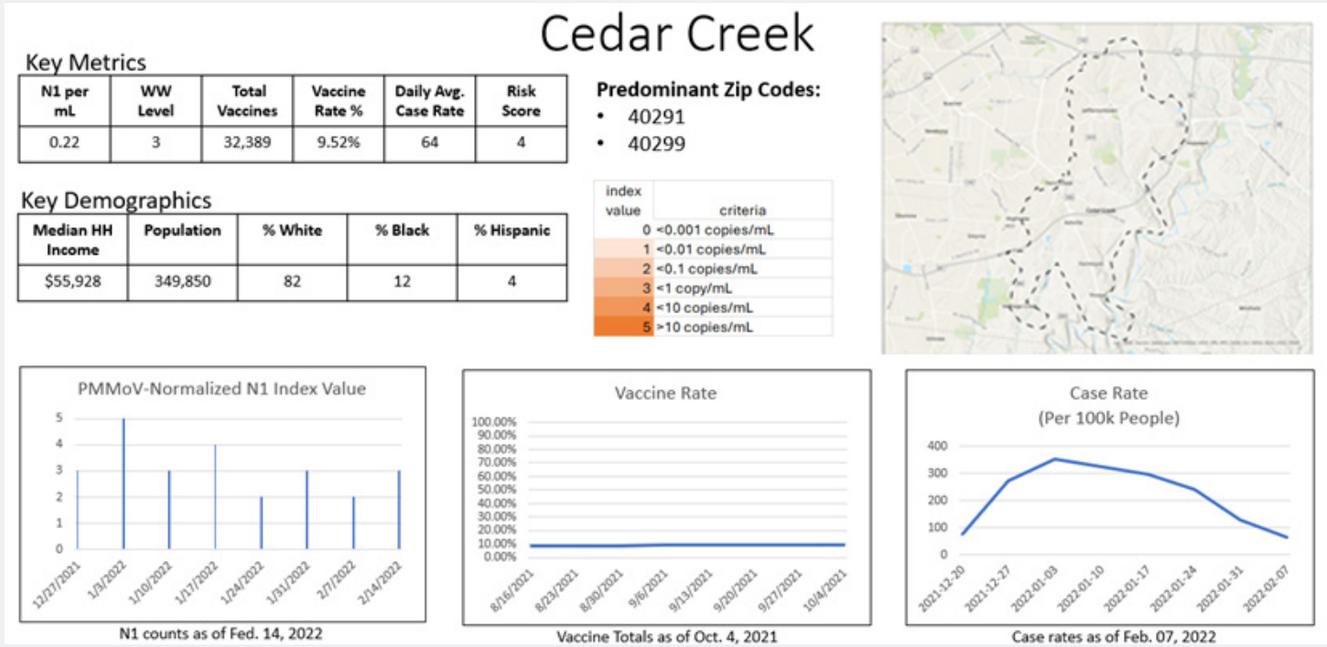


Figure 4. The weighting scheme to assign risk scores to individual neighborhoods

Wastewater Level	Wastewater Trend	Case Rate	Case Rate Trend	Vaccine Coverage
0-5/5 points based on PMMoV corrected N1 index value.	1-3/3 points based on declining, stable, or increasing N1 trend.	(0-8/8)*2 points based on current case rate.	1-3/3 points based on declining, stable, or increasing N1 trend.	1-3/3 points based on vaccine coverage.

0	0, no detect
.2	1, <10 copies/ml
.4	2, <100 copies/ml
.6	3, <1,000 copies/ml
.8	4, <10,000 copies/ml
1	5, >10,000 copies/ml

.33	Decreasing
.66	Stable
1	Increasing

0	>1
.25	1-10
.5	11-25
.75	25-37
1	38-57
1.25	58-87
1.5	88-132
1.75	133-200
2	200+

.33	Decreasing
.66	Stable
1	Increasing

.33	High (>61%)
.66	Medium (41-60%)
1	Low (<40%)