WATER CLARITY

The Need for Standardized Data Collection and Methods for Assessing and Managing Water Quality

PROJECT GOAL

We compiled a large data covering various regions, water sources, and methods. With this data, we aimed to:

(i) Asses how methodological differences affect the observed microbial water quality.

(ii) Determine if methodological factors influencing observed microbial water quality can be distinguished from non-methodological factors such as water type, weather, and land use.

Original Manuscript Information

TITLE: *Methodological differences between studies confound one-size-fits-all approaches to managing surface waterways for food and water safety*

DOI: https://doi.org/10.1128/aem.01835-23

FUNDING

- U.S. Department of Agriculture (USDA) Grant 2019-51181-30016
- Virginia Agricultural Experiment Station (VAES), Virginia Polytechnic Institute and State University

Partnering Institutions

VIRGINIA TECH

UNIVERSITY of FLORIDA

UR MEDICINE

WASHINGTON STATE UNIVERSITY

> *"Some say a microbiologist would rather use another microbiologist's toothbrush than share their methods."*

KEY RESULTS

Tidy data/Data quality

- Collected 3,211,254 data points
  - Only 2,429,990—representing 100,410 unique sites across North America—were used. (Figure 2, p. 2)
  - 781,264 samples were discarded due to data quality issues.
- Data compilation was complicated due to inconsistent and disorganized data collection, management, and reporting practices within and across organizations leading to the necessity of manually checking, verifying, and correcting each data point.
- Table 2 (p. 3) was developed using lessons learned from this study to help jump-start the standardization conversation among microbial water quality researchers and end users.

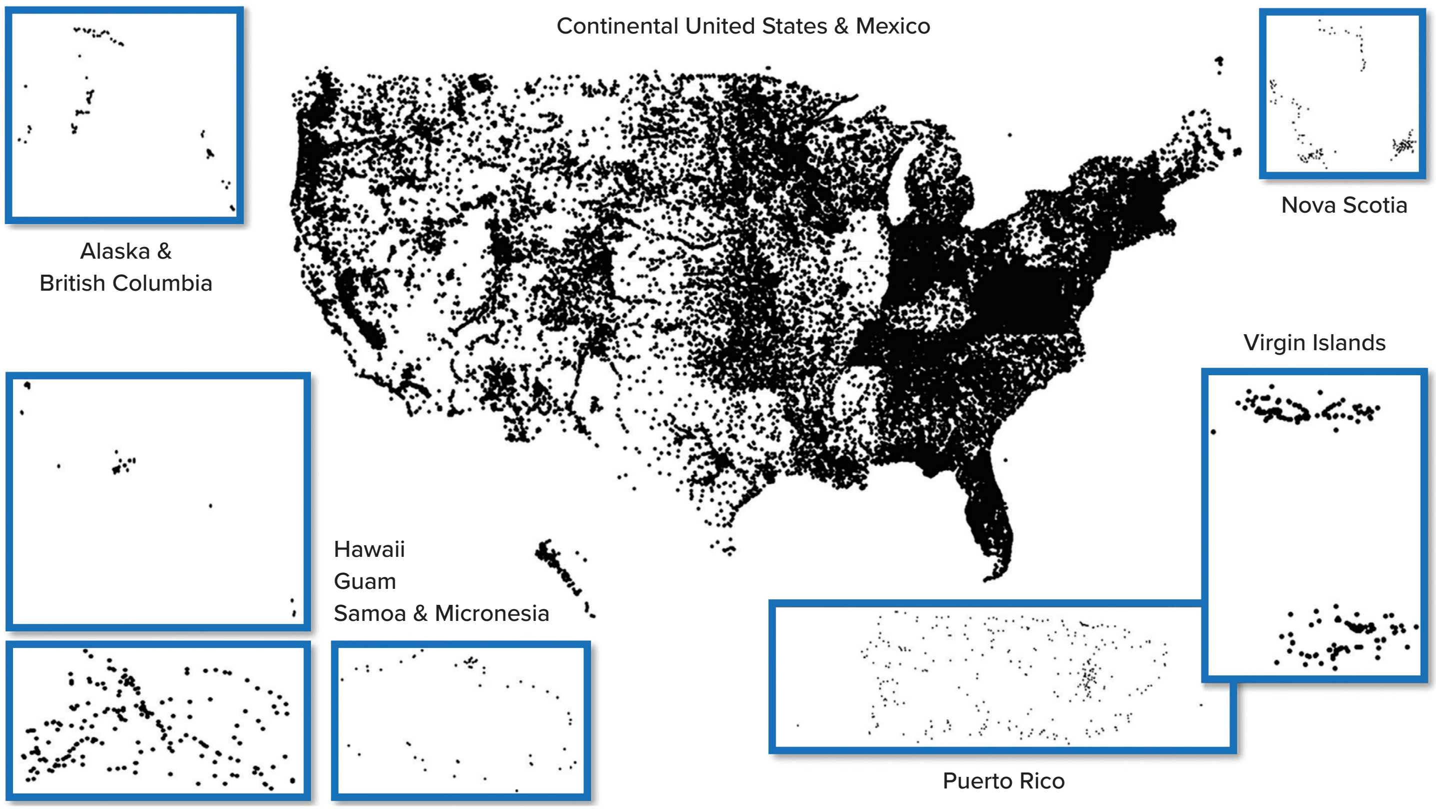

Figure 2. Data collection points from 100,410 unique sites across North America.

Methods impact observed quality

- The analysis couldn't distinguish between non-methodological (e.g., region, waterway, and water type) and methodological signals driving pathogenic *Escherichia coli*, *Salmonella*, and *Listeria* prevalence as well as indicator organism levels in water. This indicates that our understanding of the microbial ecology in water systems is confounded by variations in study methods.
- Once the waterway and site were taken into consideration, the three most influential factors linked to the probability of detecting the pathogens were:
  - *Salmonella*: filtration method, season, sample volume
  - Pathogenic *E. coli*: state, PCR gene used, culture-based vs molecular detection
  - *Listeria* spp. and *Listeria monocytogenes*: season, state, sample type (grab vs Moore swab vs modified Moore swab)
- Volume – figure 4 (p. 3):
  - Increasing sampling volume increased the odds of *Salmonella* and Pathogenic *E. coli* detection.

KEY TAKEAWAYS

- Differences in study methods confound our understanding of enteric bacteria and foodborne pathogen ecology in water systems. Without consistent methods, it's challenging to pinpoint the broader ecological factors driving contamination and to develop effective strategies for managing the associated risks.
- To enhance risk assessments and water management guidance, standardization is crucial across studies. This includes:
  - Standardizing data collection, cleaning, and management protocols.
  - Ensuring consistent reporting and archiving of water quality data, even across different organizations.
  - Establishing a minimum set of standardized data attributes for collection and reporting.
  - Standardizing sampling and laboratory methods for microbial water quality testing to enable comparability of results.

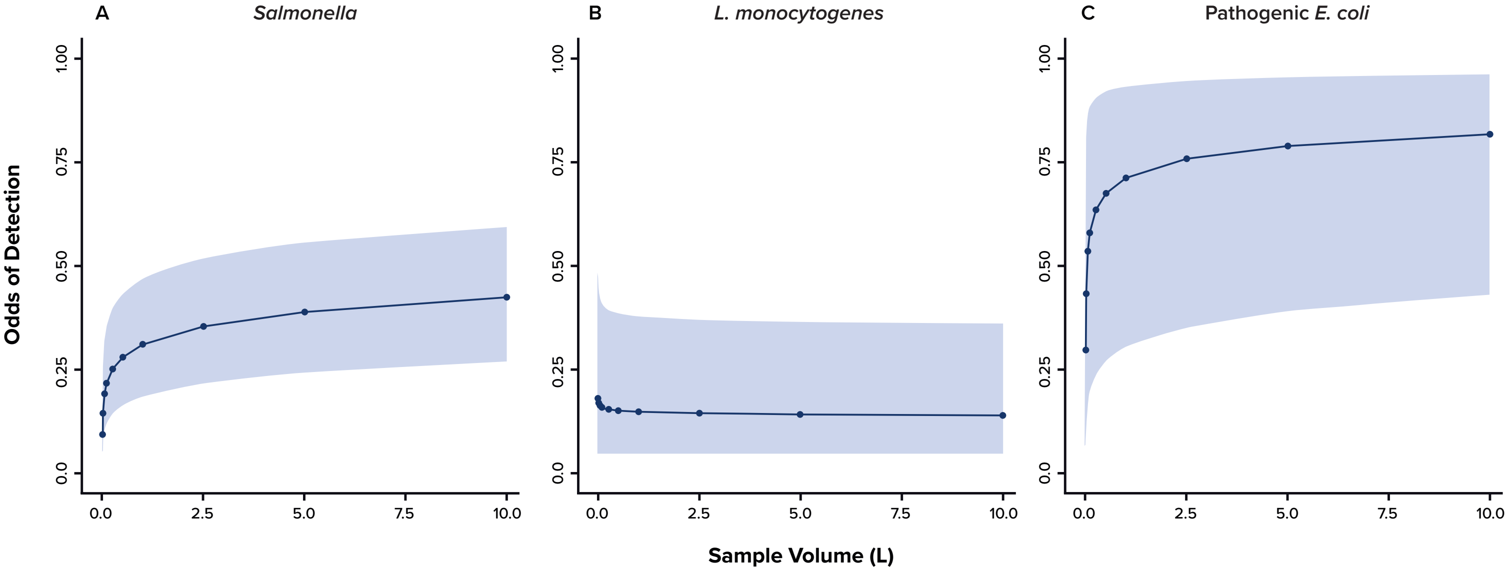

Figure 4. Increasing sampling volume increased the odds of *Salmonella* and Pathogenic *E. coli* detection.

Table 2. Recommended practices for the collection, recording, and reporting of data and attributes for research collecting water samples aimed at understanding microbial water quality.

CATEGORY	BEST PRACTICES FOR COLLECTION, RECORDING, AND REPORTING	RESOURCES & REFERENCES
Tidy Data[a]		
Data Structure	Each column represents a unique variable with a unique column header that avoids spaces, capital letters, or special characters (e.g., water_type or watertype).	(131, 132)
	Each row represents a unique observation with a unique row id.	
	Each cell is a single measurement without units within the cell (include units in data dictionary and, if needed, create a new column for units).	
	Avoid visual formatting (e.g., colored cells, borders).	
Value Formatting	For categorical variables use dropdown menus when possible so you are only selecting from pre-standardized categories. If drop-down menus aren't possible, ensure consistency in value formatting and the case of the text (e.g., avoid using "Pond", "pond", and "P" to all refer to pond samples). Ensure consistency in spelling and using of white space/special characters (e.g., avoid "pond").	
	For numeric, set upper and lower bounds so you can catch entry or measurement errors.	
Missing Values	Encode missing values as NA or as a blank. This will ensure it is read as missing data by analysis programs. Do not encode missing value as a number (e.g., 0 or 999), character (e.g., -), or word (e.g., Missing).	
Data Dictionary[b]		
	Include the column header exactly as seen in the dataset.	(133–135)
	Define each variable, including all possible entries/factor levels and their meaning, range of possible numeric values, or accepted values for the variable (e.g., ≥0, ≤ 100), and units (when applicable). If a numeric variable has upper and lower limits of detection, report these and how they should be dealt with for analysis.	
	If any imputation or data transformations were or should be employed prior to use describe these.	
	If the variable can be used alone or in conjunction with other columns to calculate new columns, explain this as well.	

[a] Tidy data provide a standard way to organize data values within a dataset that has been cleaned in a way that is ready for analysis.

[b] A data dictionary is a centralized repository of attributes that provides a comprehensive description of the data used. Its main purpose is to provide additional context and information about each data point so that analysts can understand the data better.

Table continued

Table 2. Continued.

Recommended Minimum Collected and Recorded Attributes		
Methodology		(136–140)
Sample Type	Examples include: grab sample, Moore swab	
Sample Volume	Water volume tested for the given target, NOT the volume of the sample; these may be the same but often the sample is divided into aliquots used to test for different targets. Ensure all data are entered in the same units (e.g., MPN/100mL, CFU/mL).	
Sample Site	Ideally this would be at least three columns: the sampling site latitude, the sampling site longitude, and a descriptive column (e.g., near uptake pump; 1m from into pond from pump). Additional columns that are recommended are sample depth and distance from shore.	
Unique Sampler ID	The person who collected the sample, if multiple individuals have the same initials do not use initials for this column, use an unambiguous method for this ID.	
Detection or Qualification Method	If only one method is used throughout the study clearly state the method used for target detection or enumeration in the data dictionary. The description should provide sufficient detail [e.g., was it a culture-based or molecular method, what was the detection limit(s), if and how the sample was filtered (e.g., membrane, none, modified Moore swab), if and how the target was enumerated (e.g., most probable number-based approach, membrane filter-based), volume used for reporting results (e.g., count per 100 mL, count per 10 ml)].	
	If there is a standard name for the method (e.g., EPA Method 1603) include that. Do not use organization or lab-specific names in this column, as this information will not mean anything to folks outside the organization or lab.	
	Be clear what the target is (e.g., is it the microbe, a specific gene, multiple genes).	
	If the method has previously been validated/published in the peer-reviewed research literature, include this reference. Even if it was published include key performance information for the method (e.g., sensitivity, specificity).	
	If positive and negative controls were used, include what those controls were.	
Spatial		(133, 140–143)
GPS Coordinates	Use a consistent coordinate reference system (i.e., DATUM) and note this in the data dictionary.	
	Use a standard format; formats that include spaces, multiple symbols, or multiple “.” can result in coordinates reading in incorrectly. Confirm the appropriate presence/absence of “-” if that is part of the format you use. DO NOT drop the “-” just because all of your sites are in the same hemisphere.	
	Include longitude and latitude as separate columns.	
Hydrologic Units (HUC)	6 or 9-digit code	
Type of waterway	Examples include: pond, stream, canal	
Location	This could include separate columns for county, state, county, and/or city.	
Unique Site ID	Make it unique to each site and unambiguous. Use a standard way of naming that is relevant to your study design.	
Waterway Name	Common name(s) used to refer to the waterway. If multiple names, separate by “;” or”,”.	
Temporal		(138, 140, 144)
Date	MM/DD/YYYY	
Time of Day	Use military time (24 hours) to reduce the risk of incorrectly specifying am/pm.	
Physiochemical		(138, 145–147)
Turbidity **Water Temperature** **pH** **Total Suspended Solids** **Dissolved Oxygen** **Dissolved Organic Matter** **Conductivity** **Salinity**	For categorical variables, use a drop-down menu to avoid spelling or entry errors. For numeric values, use a consistent number of significant digits reflects the accuracy of measuring the device and sets minimum and maximum thresholds to catch entry/measurement errors. Use consistent units (e.g., metric, imperial) and include units in the data dictionary.	
Meteorological		(138, 147, 148)
Air Temperature **Precipitation Volume** **Relative Humidity** **UV Intensity** **Wind Speed**	For categorical variables, use a drop-down menu to avoid spelling or entry errors. For numeric values, use a consistent number of significant digits reflects the accuracy of measuring the device and sets minimum and maximum thresholds to catch entry/measurement errors. Use consistent units (e.g., metric, imperial) and include units in the data dictionary.	