AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology®

PUBLIC AND ENVIRONMENTAL HEALTH MICROBIOLOGY

Methods for Handling Left-Censored Data in Quantitative Microbial Risk Assessment

Robert A. Canales,[a] Amanda M. Wilson,[a] Jennifer I. Pearce-Walker,[a] Marc P. Verhougstraete,[a] Kelly A. Reynolds[a]

[a]Mel & Enid Zuckerman College of Public Health, The University of Arizona, Tucson, Arizona, USA

ABSTRACT Data below detection limits, left-censored data, are common in environmental microbiology, and decisions in handling censored data may have implications for quantitative microbial risk assessment (QMRA). In this paper, we utilize simulated data sets informed by real-world enterovirus water data to evaluate methods for handling left-censored data. Data sets were simulated with four censoring degrees (low [10%], medium [35%], high [65%], and severe [90%]) and one real-life censoring example (97%) and were informed by enterovirus data assuming a lognormal distribution with a limit of detection (LOD) of 2.3 genome copies/liter. For each data set, five methods for handling left-censored data were applied: (i) substitution with LOD/$\sqrt{2}$, (ii) lognormal maximum likelihood estimation (MLE) to estimate mean and standard deviation, (iii) Kaplan-Meier estimation (KM), (iv) imputation method using MLE to estimate distribution parameters (MI method 1), and (v) imputation from a uniform distribution (MI method 2). Each data set mean was used to estimate enterovirus dose and infection risk. Root mean square error (RMSE) and bias were used to compare estimated and known doses and infection risks. MI method 1 resulted in the lowest dose and infection risk RMSE and bias ranges for most censoring degrees, predicting infection risks at most 1.17×10^{-2} from known values under 97% censoring. MI method 2 was the next overall best method. For medium to severe censoring, MI method 1 may result in the least error. If unsure of the distribution, MI method 2 may be a preferred method to avoid distribution misspecification.

IMPORTANCE This study evaluates methods for handling data with low (10%) to severe (90%) left-censoring within an environmental microbiology context and demonstrates that some of these methods may be appropriate when using data containing concentrations below a limit of detection to estimate infection risks. Additionally, this study uses a skewed data set, which is an issue typically faced by environmental microbiologists.

KEYWORDS left censored, limit of detection, quantitative microbial risk assessment

Methodologies for handling data below limits of detection (LOD) have been a long-recognized issue in many scientific disciplines, and it is a frequent reality for environmental health and exposure scientists. Data below a LOD for which the true value is unknown are often referred to as "left-censored." If data are above a particular value but the true value is unknown, these are referred to as "right censored." For example, in microbiology, a left-censored data point may be one for which nothing was detected with a particular method or assay, but it is unknown whether there are truly zero organisms of interest in the sample. An example of a right-censored data point in microbiology might be a plate count of "TNTC" (too numerous to count), where the lower bound of the possible value may be known but the true value of the count is unknown.

The issue of left-censored data is familiar in environmental microbiology. Virus recovery efficiency challenges and needs for adjustments to viral concentration data to

Received 18 May 2018 **Accepted** 8 August 2018

Accepted manuscript posted online 17 August 2018

Citation Canales RA, Wilson AM, Pearce-Walker JI, Verhougstraete MP, Reynolds KA. 2018. Methods for handling left-censored data in quantitative microbial risk assessment. Appl Environ Microbiol 84:e01203-18. https://doi.org/10.1128/AEM.01203-18.

Editor Donald W. Schaffner, Rutgers, The State University of New Jersey

Address correspondence to Amanda M. Wilson, apfeifer@email.arizona.edu.

more closely reflect reality have been acknowledged (1). Additionally, more recent environmental microbiological studies have addressed that substitution is not an appropriate method for handling LOD data. In 2015, Pouillot et al. (2) used methods from survival analysis and Bayesian inference models to characterize distributions of norovirus and male-specific coliphage concentrations due to these methods' ability to handle left-censoring (2). The distributions were then used to model viral concentrations and log reductions due to various water treatment steps (2). In 2016, Vergara et al. used nondetects (left-censored values) and data analysis (NADA) methods developed by Helsel et al. in fitting distributions to norovirus and human adenovirus concentration data utilized in a quantitative microbial risk assessment (QMRA) (3, 4).

The method chosen for handling data below LOD has important implications for estimating risks within QMRA. In several sensitivity analyses for exposure models, pathogen concentration was found to be one of the biggest drivers, if not the number one driver, of infection risk (5–8). For pathogenic viruses for which there are low infectious doses relative to those of other microorganisms, the method for handling data at or below LOD may have notable impacts on predicted health outcomes. Interpreting data below the LOD within an environmental health context is particularly important, as data may be used to determine health risks associated with exposures. Substituting values below the LOD with a constant, such as the LOD, LOD/2, or LOD/$\sqrt{2}$, has been a common approach (9). However, it is well recognized that these substitution methods may introduce error (observed values − actual values), especially when large portions of a data set are below the LOD (9–11). Helsel (11) went as far as to say that manuscripts should be rejected from publication if substitution methods have been used to address data below LOD. Despite warnings against the implementation of substitution methods, they are still heavily used in some disciplines (11, 12). Among the common substitution values, LOD/$\sqrt{2}$ has been favored for lognormal distributions (10). Instead of using substitution methods, some have suggested approaches including maximum likelihood estimation (MLE) for parametric data and the Kaplan-Meier (KM) method when the underlying distribution of a data set is unknown. Some have regarded the MLE method as the "gold standard" for handling data below LOD (13). However, others have acknowledged that if there is "severe" censoring or if the data are highly skewed, then the use of a nonparametric means, such as the KM method, is recommended over the MLE method, because a distribution is not assumed in using the KM method (14). Although methods beyond substitution have been acknowledged as mathematically intensive and perhaps impractical to implement, affordability of the necessary computing power and the availability of programs or tools to allow for broader adoption have altered this reality (12).

In 2014, the U.S. EPA published a document recognizing that substitution methods may not be appropriate for microbial field data, especially considering computing advances that allow for the use of more rigorous methods, such as MLE, KM, and multiple imputation (MI) (15). However, multiple options for handling censored data were discussed (15). With many available and recommended methods existing in current and previous literature surrounding LOD issues in environmental sciences, there is a need for the exploration of these methods within a QMRA context to demonstrate the potential impact of LOD methodology on estimated risk and to identify superior methods that support accurate risk assessments. Additionally, previous simulation studies comparing LOD methods often have not incorporated highly right-skewed data, which is common in environmental microbiology (16–19). The purpose of this study was to evaluate substitution, MLE, KM, and two MI methods for lognormal highly right-skewed data sets with low, medium, high, and severe censoring, as defined by the U.S. Army Public Health Command in 2015 (14), in addition to a real-life example in which 97% of samples were left-censored (20). Left-censoring in microbial data has been shown to occur at all of these censoring levels (20–23), making these censoring degrees relevant to the field of QMRA. The effects of LOD methods applied to simulated

TABLE 1 Descriptive statistics of simulated data sets per degree of censoring category for known data sets and after implementation of censored data methods[a]

Data set or censored data method	Viral concn (genome copies/liter) (mean ± SD) for indicated degree of censoring				
	Low	Medium	High	Severe	Real-life example
Known	25.93 ± 88.35	18.87 ± 77.52	10.16 ± 56.97	2.93 ± 31.59	0.88 ± 16.66
Substitution LOD/$\sqrt{2}$	26.09 ± 88.30	19.43 ± 77.38	11.20 ± 56.79	4.38 ± 31.46	2.44 ± 16.58
MLE[b]	18.56 ± 44.31	15.51 ± 68.25	11.44 ± 225.24	$49.15 \times 10^{0} \pm 1.62 \times 10^{8}$	$1.47 \times 10^{8} \pm 2.36 \times 10^{26}$
KM[b]	26.17 ± 77.32	19.68 ± 65.06	11.70 ± 44.87	5.28 ± 19.25	5.06 ± 8.20
MI method 1	26.06 ± 88.31	**19.21 ± 77.44**	**10.54 ± 56.90**	**3.14 ± 31.57**	**1.01 ± 16.66**
MI method 2	**26.05 ± 88.32**	19.27 ± 77.42	10.90 ± 56.84	3.95 ± 31.49	1.98 ± 16.62

[a]Data sets comprised simulated data sets with low (10%), medium (35%), high (65%), and severe (90%) degrees of censoring and a real-life censoring example (97%); censored data methods included substitution LOD/$\sqrt{2}$, MLE, KM, MI method 1, and MI method 2. Bold values indicate those that were the closest to the known values for that degree of censoring.

[b]For the MLE and KM methods, where a summary statistic is estimated for the entire data set as opposed to accounting for individual values in place of censored data, standard deviations represented for all data sets at a degree of censoring are equal to the mean of the estimated standard deviations for each data set.

data on viral concentrations in drinking water on estimated viral doses and infection risks were then quantified and compared.

RESULTS

For medium, high, and severe degrees of censoring, MI method 1 (which used MLE methods to estimate parameters of the lognormal distribution and imputing censored data points with values from this distribution below the LOD) estimated mean viral concentrations and predicted infection risks closest to those of the known, unmasked data sets (Tables 1 and 2). Additionally, this method resulted in the smallest root mean square errors (RMSEs) and biases in dose and infection risk for medium to severe degrees of censoring (Tables 3 and 4). MI method 2 (which assumed that all data below the LOD followed a uniform distribution and imputed values from this distribution for censored values) was the next overall best method for estimating the mean viral concentration of the data set (Table 1). MI method 2 resulted in the lowest RMSEs and biases in dose and infection risk for low censored data (Tables 3 and 4). As degrees of censoring increased, ranges of dose and infection risk biases and RMSEs increased, meaning that performance of the methods became more variable.

Biases for low to severe censoring for MI method 1 were positive, indicating that it overpredicts risk. However, with 97% censoring, MI method 1 began to underpredict some doses and infection risks. The MLE method underpredicted some doses and infection risks for all degrees of censoring. The infection risk bias with the smallest magnitude (1.14×10^{-4}) was observed under 90% censoring using MI method 1. The bias greatest in magnitude (9.93×10^{-1}) was observed for the MLE method with 97% censoring. This means that a predicted risk value of 0.01 could be incorrectly estimated to be as large as approximately 1.00 using an MLE method with 97% censored data. The MLE method had some bias values on the order of 10^{-1} for all censoring degrees, meaning that at any censoring degree, it is possible that the MLE method could result in a predicted risk that is 0.1 larger or smaller than the true infection risk. When comparing the distributions of predicted infection risks with 90% censored data sets, the MLE method predicted a maximum infection risk of 1 where the actual maximum infection risk was 0.21 (Table 2).

In the real-life example in which 97% of data were below the LOD, MI method 1 produced the lowest dose and infection risk biases. Using this method, the largest bias for dose was 3.14 viral particles, while the smallest bias was −0.058 viral particles (Table 4). For infection risk, the largest bias measured using MI method 1 was 1.17×10^{-2}, while the smallest bias was -2.14×10^{-4} (Table 4). This means that, at worst, this method may overpredict risk by 1.17×10^{-2} or may underpredict risk by 2.14×10^{-4}.

Most of the methods, aside from the MLE method, performed well in estimating the mean, minimum, and maximum infection risks. The substitution method performed well under low, medium, and high censoring (Table 2). Under severe censoring, the mean and maximum infection risks were closely estimated, but the minimum infection risks were

TABLE 2 Descriptive statistics of predicted infection risks per degree of censoring category for known data sets and after implementation of censored data methods[a]

Data set or censored data method	Infection risk for indicated degree of censoring									
	Low		Medium		High		Severe		Real-life example	
	Mean (SD)	Min, max	Mean (SD)	Min, max	Mean (SD)	Min, max	Mean (SD)	Min, max	Mean (SD)	Min, max
Known	0.17 (0.052)	0.068, 0.38	0.13 (0.049)	0.045, 0.36	0.072 (0.036)	0.017, 0.21	0.021 (0.022)	0.0025, 0.21	0.0065 (0.011)	0.00066, 0.11
Substitution LOD/$\sqrt{2}$	**0.18 (0.052)**	0.069, 0.39	**0.13 (0.049)**	0.049, 0.36	0.079 (0.036)	0.024, 0.22	0.032 (0.022)	0.013, 0.22	0.018 (0.011)	0.012, 0.12
MLE	0.13 (0.030)	0.065, 0.27	0.11 (0.036)	0.044, 0.31	0.081 (0.048)	0.021, 0.47	0.085 (0.16)	0.0075, 1.00	0.18 (0.34)	0.0029, 1.00
KM	**0.18 (0.052)**	0.070, 0.39	0.14 (0.049)	0.051, 0.36	0.083 (0.036)	0.028, 0.22	0.038 (0.022)	0.018, 0.22	0.037 (0.028)	0.017, 0.32
MI method 1	**0.18 (0.052)**	**0.069, 0.38**	**0.13 (0.049)**	0.048, 0.36	**0.075 (0.036)**	0.028, 0.21	**0.023 (0.021)**	**0.0068, 0.21**	**0.0074 (0.011)**	**0.0021, 0.11**
MI method 2	**0.18 (0.052)**	**0.069, 0.38**	**0.13 (0.049)**	**0.047, 0.36**	0.077 (0.036)	**0.022, 0.22**	0.029 (0.022)	0.010, 0.22	0.015 (0.011)	0.0078, 0.11

[a]Data sets comprised simulated data sets with low (10%), medium (35%), high (65%), and severe (90%) degrees of censoring and a real-life censoring example (97%); censored data methods included substitution LOD/$\sqrt{2}$, MLE, KM, MI method 1, and MI method 2. Bold values indicate those that were the closest to the known mean (SD) or minimum/maximum values for that degree of censoring. Min, minimum; max, maximum.

TABLE 3 Comparison of RMSEs for predicted doses and infection risks by censoring degree and method for handling censored data[a]

Estimation and censored data method	RMSE for indicated degree of censoring									
	Low		Medium		High		Severe		Real-life example	
	Mean (SD)	Min, max	Mean (SD)	Min, max	Mean (SD)	Min, max	Mean (SD)	Min, max	Mean (SD)	Min, max
Estimated dose (viral particles)										
Substitution LOD/$\sqrt{2}$	0.32 (0.0071)	0.27, 0.33	1.13 (0.014)	1.022, 1.14	2.093 (0.018)	1.98, 2.11	2.90 (0.022)	2.80, 2.93	3.12 (0.022)	3.017, 3.15
MLE	14.75 (10.63)	0.11, 65.12	7.51 (8.32)	0.0052, 57.19	5.79 (8.74)	0.0032, 115.92	92.64 (1,105.50)	0.0040, 27,513.36	2.94×10^8 (8.64×10^9)	0.00, 2.73×10^{11}
KM	0.46 (0.010)	0.40, 0.51	1.63 (0.038)	1.51, 1.88	3.087 (0.12)	2.91, 3.76	4.70 (0.67)	4.063, 9.51	8.36 (7.20)	4.39, 94.069
MI method 1	0.25 (0.039)	0.12, 0.37	**0.68 (0.095)**	**0.40, 1.00**	**0.78 (0.14)**	**0.38, 1.32**	**0.42 (0.19)**	**0.032, 1.68**	**0.26 (0.33)**	**3.40×10^{-4}, 3.14**
MI method 2	**0.23 (0.043)**	**0.12, 0.35**	0.80 (0.078)	0.54, 1.03	1.48 (0.11)	1.15, 1.80	2.041 (0.13)	1.62, 2.44	2.20 (0.13)	1.72, 2.67
Estimated infection risk										
Substitution LOD/$\sqrt{2}$	9.99×10^{-4} (6.61×10^{-5})	7.05×10^{-4}, 1.13×10^{-3}	3.66×10^{-3} (2.13×10^{-4})	2.69×10^{-3}, 4.05×10^{-3}	7.23×10^{-3} (2.92×10^{-4})	6.14×10^{-3}, 7.72×10^{-3}	1.05×10^{-2} (2.48×10^{-4})	8.53×10^{-3}, 1.08×10^{-2}	1.15×10^{-2} (1.55×10^{-4})	1.03×10^{-2}, 1.17×10^{-2}
MLE	4.56×10^{-2} (3.09×10^{-2})	3.42×10^{-4}, 1.85×10^{-1}	2.40×10^{-2} (2.53×10^{-2})	1.82×10^{-5}, 1.54×10^{-1}	1.91×10^{-2} (2.54×10^{-2})	1.12×10^{-5}, 2.89×10^{-1}	6.43×10^{-2} (1.48×10^{-1})	1.51×10^{-4}, 9.5×10^{-1}	1.77×10^{-1} (3.4×10^{-1})	8.68×10^{-4}, 9.93×10^{-1}
KM	1.43×10^{-3} (9.46×10^{-5})	1.05×10^{-3}, 1.67×10^{-3}	5.28×10^{-3} (3.26×10^{-4})	3.86×10^{-3}, 6.41×10^{-3}	1.06×10^{-2} (5.77×10^{-4})	8.70×10^{-3}, 1.33×10^{-2}	1.70×10^{-2} (2.38×10^{-3})	1.24×10^{-2}, 3.5×10^{-2}	3.02×10^{-2} (2.39×10^{-2})	1.60×10^{-2}, 2.88×10^{-1}
MI method 1	7.78×10^{-4} (1.41×10^{-4})	3.54×10^{-4}, 1.22×10^{-3}	**2.23×10^{-3} (3.81×10^{-4})**	**1.15×10^{-3}, 3.52×10^{-3}**	**2.70×10^{-3} (5.55×10^{-4})**	**1.26×10^{-3}, 4.81×10^{-3}**	**1.55×10^{-3} (7.29×10^{-4})**	**1.14×10^{-4}, 6.91×10^{-3}**	**9.55×10^{-4} (1.22×10^{-3})**	**1.26×10^{-7}, 1.17×10^{-2}**
MI method 2	**6.94×10^{-4} (1.39×10^{-4})**	**3.46×10^{-4}, 1.1×10^{-3}**	2.59×10^{-3} (2.97×10^{-4})	1.75×10^{-3}, 3.45×10^{-3}	5.12×10^{-3} (4.25×10^{-4})	3.74×10^{-3}, 6.30×10^{-3}	7.44×10^{-3} (4.98×10^{-4})	5.79×10^{-4}, 8.98×10^{-3}	8.15×10^{-3} (4.95×10^{-4})	6.41×10^{-3}, 9.78×10^{-3}

[a]Data sets comprised simulated data sets with low (10%), medium (35%), high (65%), and severe (90%) degrees of censoring and a real-life censoring example (97%); censored data methods included substitution LOD/$\sqrt{2}$, MLE, KM, MI method 1, and MI method 2. Bold values indicate those that were the closest to the known values for that degree of censoring. Min, minimum; max, maximum.

TABLE 4 Comparison of biases for predicted doses and infection risks by censoring degree and method for handling censored data[a]

Estimation and censored data method	Bias for indicated degree of censoring									
	Low		Medium		High		Severe		Real-life example	
	Mean (SD)	Min, max	Mean (SD)	Min, max	Mean (SD)	Min, max	Mean (SD)	Min, max	Mean (SD)	Min, max
Estimated dose (viral particles)										
Substitution LOD/$\sqrt{2}$	0.32 (0.0071)	0.27, 0.33	1.13 (0.015)	1.02, 1.14	2.09 (0.018)	1.98, 2.11	2.90 (0.022)	2.80, 2.93	3.12 (0.022)	3.02, 3.15
MLE	−14.74 (10.65)	−65.12, 2.87	−6.72 (8.97)	−57.19, 19.31	2.56 (10.17)	−27.8, 115.92	92.45 (1105.51)	$-12.88, 2.75 \times 10^{-4}$	2.94×10^{8} (8.64×10^{9})	$0.00, 2.73 \times 10^{11}$
KM	0.46 (0.010)	0.40, 0.51	1.63 (0.038)	1.51, 1.88	3.09 (0.12)	2.91, 3.76	4.70 (0.67)	4.06, 9.51	8.36 (7.20)	4.39, 94.10
MI method 1	0.25 (0.039)	0.12, 0.37	**0.68 (0.095)**	**0.40, 1.00**	**0.77 (0.14)**	**0.38, 1.32**	**0.42 (0.19)**	**0.03, 1.86**	**0.25 (0.33)**	**−0.058, 3.14**
MI method 2	**0.23 (0.043)**	**0.12, 0.35**	0.80 (0.078)	0.54, 1.03	1.48 (0.11)	1.15, 1.80	2.04 (0.13)	1.62, 2.44	2.20 (0.13)	1.72, 2.67
Estimated infection risk										
Substitution LOD/$\sqrt{2}$	9.93×10^{-4} (6.61×10^{-5})	7.05×10^{-4}, 1.13×10^{-3}	3.66×10^{-3} (2.13×10^{-4})	2.69×10^{-3}, 4.05×10^{-3}	7.23×10^{-3} (2.92×10^{-4})	6.14×10^{-3}, 7.72×10^{-3}	1.05×10^{-2} (2.48×10^{-4})	8.53×10^{-3}, 1.084×10^{-2}	1.15×10^{-2} (1.55×10^{-4})	1.03×10^{-2}, 1.17×10^{-2}
MLE	-4.55×10^{-2} (3.091×10^{-2})	-1.85×10^{-1}, 9.39×10^{-3}	-2.14×10^{-2} (2.75×10^{-2})	-1.54×10^{-1}, 5.56×10^{-2}	8.11×10^{-3} (3.072×10^{-2})	-9.19×10^{-2}, 2.89×10^{-1}	6.37×10^{-2} (1.49×10^{-1})	-4.41×10^{-2}, 9.46×10^{-1}	1.77×10^{-1} (3.36×10^{-1})	8.68×10^{-4}, 9.93×10^{-1}
KM	1.43×10^{-3} (9.46×10^{-5})	1.05×10^{-3}, 1.69×10^{-3}	5.28×10^{-3} (3.26×10^{-4})	3.86×10^{-3}, 6.41×10^{-3}	1.065×10^{-2} (5.77×10^{-4})	8.70×10^{-3}, 1.33×10^{-2}	1.70×10^{-2} (2.38×10^{-3})	1.24×10^{-2}, 3.46×10^{-2}	3.02×10^{-2} (2.39×10^{-2})	1.60×10^{-2}, 2.88×10^{-1}
MI method 1	7.78×10^{-4} (1.41×10^{-4})	3.54×10^{-4}, 1.22×10^{-3}	**2.23×10^{-3} (3.81×10^{-4})**	**1.15×10^{-3}, 3.52×10^{-3}**	**2.68×10^{-3} (5.54×10^{-4})**	**1.26×10^{-3}, 4.81×10^{-3}**	**1.55×10^{-3} (7.29×10^{-4})**	**1.14×10^{-4}, 6.91×10^{-3}**	**9.48×10^{-4} (1.23×10^{-3})**	**-2.14×10^{-4}, 1.17×10^{-2}**
MI method 2	**6.99×10^{-4} (1.40×10^{-4})**	**2.50×10^{-4}, 1.10×10^{-3}**	2.58×10^{-3} (2.84×10^{-4})	1.70×10^{-3}, 3.51×10^{-3}	5.10×10^{-3} (4.22×10^{-4})	3.21×10^{-3}, 6.55×10^{-3}	7.44×10^{-3} (4.98×10^{-4})	5.79×10^{-3}, 8.98×10^{-3}	8.16×10^{-3} (4.95×10^{-4})	6.41×10^{-3}, 9.78×10^{-3}

[a]Data sets comprised simulated data sets with low (10%), medium (35%), high (65%), and severe (90%) degrees of censoring and a real-life censoring example (97%); censored data methods included substitution LOD/$\sqrt{2}$, MLE, KM, MI method 1, and MI method 2. Bold values indicate those that were the closest to the known values for that degree of censoring. Min, minimum; max, maximum.

overestimated. A similar pattern was observed for KM and MI method 2 (Table 2). MI method 1 consistently estimated an infection risk that was, at most, 1.17×10^{-2} from the actual value (Table 4), and minimum and maximum infection risks were closely predicted, even when the level of censoring was severe (Table 2).

All infection risks resulting from uncensored simulated data sets resulted in risks greater than the risk target, 1/10,000 (Table 2). This demonstrates that even when the 50th percentile of a distribution would result in a risk target, using the mean concentration of highly right-skewed data sets could result in risk predictions roughly 1- to 2-$\log_{10}$ larger than the risk target.

The sensitivity analysis demonstrated that 25% increases or decreases in the assumed geometric mean, geometric standard deviation, and LOD used to inform the assumed distribution for simulated data sets did not affect the conclusions that MI method 1 was overall superior in handling medium, high, and severe degrees of censoring and the real-life example of censoring while MI method 2 was superior in handling low degrees of censoring (see Tables S1 to S4 in the supplemental material).

DISCUSSION

In a comparison of all measures of performance, MI method 1 was overall the most superior method, because it resulted in the smallest biases and RMSEs and captured distributions of infection risk well for the largest number of censoring categories (medium, high, and severe censoring) and for the real-life example. Aside from being useful in providing accurate estimations of censored data, this method slightly overpredicted the true risk for most censoring levels, making it a conservative tool in assessing risk and preferred over a method that may underestimate the true risk.

MI method 2 consistently performed the best for low censoring. However, it is possible that other methods may have been superior if a different parameter estimate, aside from the mean, had been used as a point of comparison. Within a QMRA context, means are typically used in deterministic modeling because this parameter is more conservative than the median, as it is sensitive to high exposure concentrations that may be experienced (16, 18).

The MLE method did not perform as well as other methods in these simulations. Although the MLE method has performed well in other simulation studies, other work has noted that the MLE method does not perform as well for highly skewed data, producing larger mean square errors (24). The substitution method did not perform as poorly as expected, as this method outperformed the MLE and KM methods. However, because R statistical software packages exist that make the implementation of multiple-imputation methods accessible, substitution methods should not be the first approach in handling left-censored data (25).

This study did not consider a scenario in which distribution misspecification occurs. If the assumed distribution is not the distribution by which the true data abide, this could lead to poor performance of methods that assume a particular distribution (24). If one is unsure of the distribution of the data for data sets with medium to severe degrees of censoring, MI method 2 may be more reliable than MI method 1, as distribution misspecification could result in poor performance of MI method 1. However, if one is confident that the distribution of the data has been accurately identified or if the assumption is strongly informed, then MI method 1 may be more appropriate to use. A source of uncertainty in this study was assumptions regarding microbial distribution characteristics that informed simulated data sets. Although real-life values from a microbial data set were used, future studies should evaluate these methods for other anticipated or observed microbial data set distributions.

Both MI methods are easily applicable within QMRA, as opposed to the MLE or KM methods, because they allow for the use of the full data set and are not restricted to using summary statistics for the data set. Use of the full data set is helpful if the intention is to utilize it in stochastic exposure modeling. MI method 1 could be used to fit distributions to data sets containing left-censored data so that these distributions could later be randomly sampled to represent a variety of expected concentrations. MI

method 2 could be used to create a mixed distribution, where the uniform distribution would be sampled for a certain percentage of the time to represent the proportion of censored data in the data set. The uncensored data could then be sampled from discretely, or one could sample from a distribution fit to these uncensored concentrations (26). This study identifies methods for handling highly skewed, left-censored microbial concentration data and quantifies the impact of various methods on accuracy in calculating summary statistics and estimating infection risks with concentration data. Adoption of more mathematically intensive methods that offer better accuracy in handling left-censored data, such as the MI methods suggested in this study, are becoming more practical, as free software packages are available. Methods evaluated in this study can be used to predict infection risks associated with quantified microbial concentrations. There are a number of previous microbiological and QMRA studies that could have benefited from these methods in handling left-censored data due to various levels of left-censoring (21, 27, 28). In a study conducted by Herzog et al., the LOD was one of the most sensitive factors for health risk estimation in multiple modeled scenarios (29). As efforts continue to develop or modify current microbial detection methods with greater sensitivity, statistical methods, such as the ones in this study, can be utilized to bridge the current gap in addressing commonly experienced LOD issues in microbiology.

MATERIALS AND METHODS

Simulating quantitative PCR drinking water virus concentration data. R statistical software was used for all simulations and evaluations in this study (25). A distribution that would be used to simulate the known data sets was first created by calculating an enterovirus concentration (genome copies/liter) that would be needed to result in an annual 1/10,000 infection risk, based on assumed parameters for the equations used to estimate dose and infection risk.

A lognormal distribution for virus concentrations was assumed, as this distribution has been applied to microbial water quality data, specifically for enteric viruses, and for other types of environmental concentration data (1, 21, 30). The concentration that would result in a 1/10,000 annual infection risk, the tolerable risk established by the U.S. EPA, using the following equations was set equal to the geometric mean of the lognormal distribution (3.66×10^{-5} genome copies/liter) used to create the "known" simulated data set (31). The geometric standard deviation of detected enterovirus in water samples (78.2 genome copies/liter) from a study conducted by Pearce-Walker in 2017 was used to represent the expected variability of viral concentrations in drinking water samples (20). The LOD for simulated data sets (2.3 genome copies/liter) was calculated using methods described by Pearce-Walker, assuming a filtered sample volume of 1,000 liters. Although not the processed sample volume utilized by Pearce-Walker (20), we used this larger sample volume, sometimes used when sampling water samples for viruses (32), in order to create a scenario in which LOD may be relatively close to infectious doses of viruses. In the case of this scenario, the LOD (2.3 genome copies/liter) was 2 orders of magnitude lower than the enterovirus infectious dose estimated by the QMRA wiki (http://qmrawiki.canr.msu.edu/index.php/Quantitative_Microbial_Risk_Assessment_(QMRA)_Wiki). However, some viruses, such as poliovirus and rotavirus, have infectious doses estimated to be as low as 1.41 and 6.17 viral particles, respectively (http://qmrawiki.canr.msu.edu/index.php/Quantitative_Microbial_Risk_Assessment_(QMRA)_Wiki). In this study, it was assumed that all detected viruses were viable, as this has been a recommended conservative approach in other enteric virus QMRA contexts (33). Equation 1 was used to estimate enterovirus exposures, equivalent to dose in this case, and has been used by the World Health Organization for microbial drinking water exposure estimates (34):

$$\text{dose} = V \cdot C \tag{1}$$

where dose is the number of infectious, viable virus particles, V is the volume of water consumed per day (liters), and C is the virus concentration (viral particles/liter).

It was assumed that a person drinks 2 liters of water per day, as this value has been recommended in other drinking water QMRA contexts (35). To represent an organism with a low infectious dose, the exponential dose-response model for enterovirus was used, as this has been recommended by the QMRA wiki (http://qmrawiki.canr.msu.edu/index.php/Quantitative_Microbial_Risk_Assessment_(QMRA)_Wiki):

$$P_{\text{infect, daily}} = 1 - e^{-k \cdot \text{dose}} \tag{2}$$

where $P_{\text{infect, daily}}$ is the daily infection risk from drinking water and k equals 3.74×10^{-3}, a constant recommended by the QMRA wiki (http://qmrawiki.canr.msu.edu/index.php/Quantitative_Microbial_Risk_Assessment_(QMRA)_Wiki). The parameter k represents the probability of a single organism surviving and infecting the host (36). The annual infection risk was estimated as follows:

$$P_{\text{infect, annual}} = 1 - (1 - P_{\text{infect, daily}})^{365} \tag{3}$$

where $P_{\text{infect, annual}}$ is the annual infection risk.

Four degrees of censoring—low (10%), medium (35%), high (65%), and severe (90%)—within defined ranges stated by the U.S. Army Public Health Command (14) were considered. Additionally, 97%

censoring was investigated as a "real-life example," informed by an enterovirus data set in which approximately >96% (183/190) samples were censored (20). For each degree of censoring, 1,000 simulated data sets, assumed as "true" known data for our purposes, were created. Each individual data set included 100 viral concentrations, and this data set was then saved such that all data points were known, including concentrations below the theoretical LOD. A copy of this data set was then altered so that either 10%, 35%, 65%, 90%, or 97% of the concentrations were below the theoretical LOD concentration assumed in this study. Methods for handling left-censored data were applied to these censored data sets, and outcomes were compared to our "true" outcomes.

Substitution methods. Although multiple substitution values (0, LOD/$\sqrt{2}$, LOD/2, LOD) have been used for replacing data below LOD, LOD/$\sqrt{2}$ was utilized as the substitution method in this study, as it has been recommended over other substitution methods (10, 37). Although it has been recognized that substitution methods are only appropriate, if at all, with low-degree censoring data, this method was used on all-degree censoring data sets to demonstrate how misuse of LOD methods may impact QMRA results and to evaluate its performance for highly skewed data (9, 10).

Maximum likelihood estimation and Kaplan-Meier methods. Using the NADA package in R, MLE and KM methods were used. The NADA package uses methods by Helsel (25, 38, 39). In using the cenmle and cenfit functions, inputted data were labeled as censored or uncensored. For censored values, the LOD was used as a placeholder for these values. As MLE and KM are not imputation methods, censored values were not replaced with a value. Rather, summary statistics were estimated for the entire data set, including censored concentrations. More information regarding the MLE and KM methods implemented by the NADA package can be found at https://cran.r-project.org/web/packages/NADA/NADA.pdf.

Multiple-imputation methods. Two distribution-based multiple-imputation methods were used in this study. This method involves assuming that the entire data set, including values that fall below the LOD, follows a particular distribution. This distribution is then used to impute values for censored data. This approach has been used in other left-censoring methodology studies, and its use within an environmental context has been encouraged (9).

The first multiple-imputation method (MI method 1) used MLE methods to estimate the parameters of a lognormal distribution fit to the full simulated data set, including censored concentrations. Values lower than the LOD were then imputed from this distribution for all censored values. To estimate the parameters of the lognormal distribution, the function fitdistcens from the R package fitdistRplus was used (40). This method has performed well in other simulation studies addressing environmental censored data (9).

The second multiple-imputation method (MI method 2) assumed a uniform distribution (minimum = 0, maximum = LOD) for all values less than the LOD. Left-censored values were then replaced with a number randomly selected from this uniform distribution (26).

Comparing estimated doses. RMSEs were calculated to compare estimated doses and infection risks with known values, where a lower RMSE value indicates closer estimation to the known value. This method has been utilized in other studies to evaluate methods for handling left-censored data (41, 42). Biases were also calculated to evaluate the direction of error for each LOD method. A smaller magnitude of bias indicated a closer estimation to the true value.

Sensitivity analysis. To address uncertainty in the geometric mean, geometric standard deviation, and LOD used to define the distribution for creating simulated data sets, a sensitivity analysis was conducted. The geometric mean, geometric standard deviation, and LOD were individually decreased and increased by 25%. The dose and infection risk biases and RMSEs for each method were calculated and compared to baseline values.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/AEM.01203-18.

SUPPLEMENTAL FILE 1, PDF file, 0.3 MB.

ACKNOWLEDGMENTS

A.M.W. was supported by a Mel and Enid Zuckerman College of Public Health award and by the Western Alliance to Expand Student Opportunities (WAESO) Louis Stokes Alliance for Minority Participation (LSAMP) Bridge to Doctorate (BD) National Science Foundation (NSF) grant no. HRD-1608928.

REFERENCES

1. Petterson S, Grøndahl-Rosado R, Nilsen V, Myrmel M, Robertson LJ. 2015. Variability in the recovery of a virus concentration procedure in water: implications for QMRA. Water Res 87:79–86. https://doi.org/10.1016/j.watres.2015.09.006.
2. Pouillot R, van Doren JM, Woods J, Plante D, Smith M, Goblick G, Roberts C, Locas A, Hajen W, Stobo J, White J, Holtzman J, Buenaventura E, Burkhardt W, Catford A, Edwards R, DePaola A, Calci KR. 2015. Meta-analysis of the reduction of norovirus and male-specific coliphage concentrations in wastewater treatment plants. Appl Environ Microbiol 81:4669–4681. https://doi.org/10.1128/AEM.00509-15.
3. Helsel DR. 2012. Statistics for censored environmental data using Minitab and R. John Wiley and Sons, Hoboken, NJ.
4. Vergara GGRV, Rose JB, Gin KYH. 2016. Risk assessment of noroviruses and human adenoviruses in recreational surface waters. Water Res 103:276–282. https://doi.org/10.1016/j.watres.2016.07.048.
5. Julian TR, Canales RA, Leckie JO, Boehm AB. 2009. A model of exposure

to rotavirus from nondietary ingestion iterated by simulated intermittent contacts. Risk Anal 29:617–632. https://doi.org/10.1111/j.1539-6924.2008.01193.x.
6. Beamer PI, Plotkin KR, Gerba CP, Sifuentes LY, Koenig DW, Reynolds KA. 2015. Modeling of human viruses on hands and risk of infection in an office workplace using micro-activity data. J Occup Environ Hyg 12:266–275. https://doi.org/10.1080/15459624.2014.974808.
7. Blokker M, Smeets P, Medema G. 2014. QMRA in the drinking water distribution system. Procedia Eng 89:151–159. https://doi.org/10.1016/j.proeng.2014.11.171.
8. Amha YM, Kumaraswamy R, Ahmad F. 2015. A probabilistic QMRA of Salmonella in direct agricultural reuse of treated municipal wastewater. Water Sci Technol 71:1203–1211. https://doi.org/10.2166/wst.2015.093.
9. Chen H, Quandt SA, Grzywacz JG, Arcury TA. 2013. A Bayesian multiple imputation method for handling longitudinal pesticide data with values below the limit of detection. Environmetrics 24:132–142. https://doi.org/10.1002/env.2193.
10. Hornung RW, Reed LD. 1990. Estimation of average concentrations in the presence of nondetectable values. Appl Occup Environ Hyg 5:46–51. https://doi.org/10.1080/1047322X.1990.10389587.
11. Helsel D. 2010. Much ado about next to nothing: incorporating nondetects in science. Ann Occup Hyg 54:257–262.
12. Finkelstein MM, Verma DK. 2001. Exposure estimation in the presence of nondetectable values: another look. Am Ind Hyg Assoc J 62:195–198.
13. Ganser GH, Hewett P. 2010. An accurate substitution method for analyzing censored data. J Occup Environ Hyg 7:233–244. https://doi.org/10.1080/15459621003609713.
14. U.S. Army Public Health Command. 2015. How to handle censored industrial hygiene data. Technical information paper no. 55-039-0615. U.S. Army Public Health Command, Aberdeen Proving Ground, MD. https://phc.amedd.army.mil/PHC%20Resource%20Library/HowtoHandleCensoredIndustrialHygieneData_TIP_No_55-039-0615.pdf. Accessed 25 June 2018.
15. U.S. Environmental Protection Agency. 2014. Evaluation of options for interpreting environmental microbiology field data results having low spore counts. Report no. EPA/600/R-14/331. U.S. Environmental Protection Agency, Cincinnati, OH.
16. Benke KK, Hamilton AJ. 2008. Quantitative microbial risk assessment: uncertainty and measures of central tendency for skewed distributions. Stoch Environ Res Risk Assess 22:533–539. https://doi.org/10.1007/s00477-007-0171-9.
17. El-Shaarawi AH, Esterby SR, Dutka BJ. 1981. Bacterial density in water determined by Poisson or negative binomial distributions. Appl Environ Microbiol 41:107–116.
18. Haas CN. 1996. How to average microbial densities to characterize risk. Water Res 30:1036–1038. https://doi.org/10.1016/0043-1354(95)00228-6.
19. Teunis PFM, Medema GJ, Kruidenier L, Havelaar AH. 1997. Assessment of the risk of infection by Cryptosporidium or Giardia in drinking water from a surface water source. Water Res 31:1333–1346. https://doi.org/10.1016/S0043-1354(96)00387-9.
20. Pearce-Walker JI. 2017. Evaluation of human and cattle viruses as indicators of fecal contamination in irrigation water. University of Arizona, Tucson, AZ.
21. Kundu A, McBride G, Wuertz S. 2013. Adenovirus-associated health risks for recreational activities in a multi-use coastal watershed based on site-specific quantitative microbial risk assessment. Water Res 47:6309–6325. https://doi.org/10.1016/j.watres.2013.08.002.
22. Søborg DA, Hendriksen B, Kilian M, Kroer N. 2013. Widespread occurrence of bacterial human virulence determinants in soil and freshwater environments. Appl Environ Microbiol 79:5488–5497. https://doi.org/10.1128/AEM.01633-13.
23. Sassoubre LM, Love DC, Silverman AI, Nelson KL, Boehm AB. 2012. Comparison of enterovirus and adenovirus concentration and enumeration methods in seawater from Southern California, USA and Baja Malibu, Mexico. J Water Health 10:419–430. https://doi.org/10.2166/wh.2012.011.
24. Shoari N, Dubé JS, Chenouri S. 2015. Estimating the mean and standard deviation of environmental data with below detection limit observations: considering highly skewed data and model misspecification. Chemosphere 138:599–608. https://doi.org/10.1016/j.chemosphere.2015.07.009.
25. R Core Team. 2017. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
26. Canales RA. 2004. The cumulative and aggregate simulation of exposure framework. PhD dissertation, Stanford University, Stanford, CA. https://search.proquest.com/docview/305124116/abstr/B5ACF3419D014E29PQ/1?accountid=8360. Accessed 25 June 2018.
27. Ganime AC, Carvalho-Costa FA, Mendonça MCL, Vieira CB, Santos M, Filho RC, Miagostovich MP, Paulo J, Leite G. 2012. Group A rotavirus detection on environmental surfaces in a hospital intensive care unit. Am J Infect Control 40:544–547. https://doi.org/10.1016/j.ajic.2011.07.017.
28. Soller JA, Eftim S, Wade TJ, Ichida AM, Clancy JL, Johnson TB, Schwab K, Ramirez-Toro G, Nappier S, Ravenscroft JE. 2016. Use of quantitative microbial risk assessment to improve interpretation of a recreational water epidemiological study. Microb Risk Anal 1:2–11. https://doi.org/10.1016/j.mran.2015.04.001.
29. Herzog AB, McLennan SD, Pandey AK, Gerba CP, Haas CN, Rose JB, Hashsham SA. 2009. Implications of limits of detection of various methods for Bacillus anthracis in computing risks to human health. Appl Environ Microbiol 75:6331–6339. https://doi.org/10.1128/AEM.00288-09.
30. Ito T, Kato T, Takagishi K, Okabe S, Sano D. 2015. Bayesian modeling of virus removal efficiency in wastewater treatment processes. Water Sci Technol 72:1789–1795. https://doi.org/10.2166/wst.2015.402.
31. Lechevallier M, Buckley M. 2007. Clean water—what is acceptable microbial risk? American Academy of Microbiology colloquium report. American Society for Microbiology, Washington, DC. https://www.asm.org/index.php/colloquium-reports/item/4468-clean-water-what-is-acceptable-microbial-risk. Accessed 13 July 2018.
32. Ikner LA, Gerba CP, Bright KR. 2012. Concentration and recovery of viruses from water: a comprehensive review. Food Environ Virol 4:41–67. https://doi.org/10.1007/s12560-012-9080-2.
33. Van Abel N, Schoen ME, Kissel JC, Meschke JS. 2017. Comparison of risk predicted by multiple norovirus dose-response models and implications for quantitative microbial risk assessment. Risk Anal 37:245–264. https://doi.org/10.1111/risa.12616.
34. World Health Organization. 2011. Guidelines for drinking-water quality, 4th ed. WHO, Geneva, Switzerland. http://www.who.int/water_sanitation_health/publications/2011/dwq_guidelines/en/.
35. Schijven J, Derx J, de Roda Husman AM, Blaschke AP, Farnleitner AH. 2015. QMRAcatch: microbial quality simulation of water resources including infection risk assessment. J Environ Qual 44:1491. https://doi.org/10.2134/jeq2015.01.0048.
36. Jones RM, Su Y. 2015. Dose-response models for selected respiratory infectious agents: Bordetella pertussis, group a Streptococcus, rhinovirus and respiratory syncytial virus. BMC Infect Dis 15:90. https://doi.org/10.1186/s12879-015-0832-0.
37. Croghan CW, Egeghy PP. 2003. Methods of dealing with values below the limit of detection using SAS. Presented at the Southeastern SAS User Group, St. Petersburg, FL, 22 to 24 September 2003.
38. Lee L. 2017. NADA: nondetects and data analysis for environmental data R package version 1.6-1. https://CRAN.R-project.org/package=NADA.
39. Helsel DR. 2005. Nondetects and data analysis: statistics for censored environmental data. John Wiley and Sons, Hoboken, NJ.
40. Delignette-Muller ML, Dutang C. 2015. fitdistrplus: an R package for fitting distributions. J Stat Softw 64:1–34. https://doi.org/10.18637/jss.v064.i04.
41. Helsel DR, Gilliom RJ. 1986. Estimation of distributional parameters for censored trace level water quatility data. 2. Verification and application. Water Resour Res 22:147–155. https://doi.org/10.1029/WR022i002p00147.
42. Helsel DR. 1990. Less than obvious: statistical treatment of data below the detection limit. Environ Sci Technol 24:1766–1774. https://doi.org/10.1021/es00082a001.