Environmental Science
Processes & Impacts

PAPER

View Article Online
View Journal | View Issue

Cite this: *Environ. Sci.: Processes Impacts*, 2019, **21**, 893

Comparison of analytical techniques to explain variability in stored drinking water quality and microbial hand contamination of female caregivers in Tanzania†

Angela R. Harris, *[ab] Amy J. Pickering,[ac] Alexandria B. Boehm,[a] Mwifadhi Mrisho[d] and Jennifer Davis[ae]

Exposure to fecal contamination continues to be a major public health concern for low-income households in sub-Saharan Africa. Drinking water and hands are known transmission routes for pathogens in household environments. In an effort to identify explanatory variables of water and hand contamination, a variety of analytical approaches have been employed that model variation in *E. coli* contamination as a function of behaviors and household characteristics. Using data collected from 1217 households in Bagamoyo, Tanzania, this investigation compares the explanatory variables identified in the three different modeling methods to explain hand and water contamination: ordinary least squares regression, logistic regression, and classification tree. Although the modeling approaches varied, there were some similarities in the results, with certain explanatory variables being consistently identified as being related to hand and water contamination (*e.g.*, water source type for the water models and activity prior to sampling for the hand models). At the same time, there were also marked differences across the models. In sum, these results suggest there are benefits to using multiple analysis methods to assess relationships in complex systems. The models were also characterized by low explanatory power, suggesting that variation in hand and water contamination is difficult to capture when analyzing one-time water and hand rinse samples. For improved model performance, future studies could explore modeling of repeat measures of water quality and hand contamination.

Received 10th October 2018
Accepted 29th March 2019

DOI: 10.1039/c8em00460a

rsc.li/espi

Environmental significance

Assessment of microbiological contamination found in stored drinking water and on hands has been conducted extensively as a part of research investigations and monitoring efforts. Researchers seek to identify covariates for explaining contamination on these sources of fecal contamination exposure in low-income countries. Data analysis is typically conducted using bivariate tests, or sometimes, more complex regression models, but with limited success in explaining variation. This study uses data collected from ~1200 households in Tanzania to explore the use of multiple analytical techniques to explain hand and water contamination. Although approaches varied, there were similarities in the results, with certain covariates being consistently identified. The analysis highlights limitations with current practices of microbial sampling and analysis and suggests further research.

1. Introduction

Exposure to water contaminated with feces is a major public health concern in sub-Saharan Africa (SSA), where only 24% of the population has water piped into their home or yard.[1] Instead, most SSA households rely on shared point sources from which they collect water, then store it in the home for use throughout the day.[1] Such sources can provide water of reasonably good microbiological quality at the point of collection. Numerous investigations have demonstrated, however, that water quality often deteriorates during transport and storage, such that water supplies at the point of use are highly contaminated.[2] Some studies conclude that hands entering stored water could be a major contamination source.[3–7] Others claim that hands themselves serve as a transmission pathway in the fecal-oral route of diarrheal disease, independent of the water pathway.[8]

[a]*Environmental and Water Studies, Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, USA*
[b]*Civil, Construction and Environmental Engineering, North Carolina State University, 2501 Stinson Drive/208 Mann Hall, Campus Box 7908, Raleigh, NC 27695-7908, USA. E-mail: aharris5@ncsu.edu; Fax: +1 919 515 7908; Tel: +1 919 515 2402*
[c]*Civil and Environmental Engineering, Tufts University, Medford, MA, USA*
[d]*Ifakara Health Institute, Bagamoyo, Tanzania*
[e]*Woods Institute for the Environment, Stanford University, Stanford, CA, USA*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c8em00460a

In an effort to inform the design of interventions that minimize the levels of fecal contamination found in water and on hands, researchers have sought to identify household characteristics and practices associated with microbiological contamination levels. Some have employed bivariate statistical tests, evaluating the association between the concentration of fecal indicator bacteria (such as *E. coli*) and household characteristics and practices one at a time.[9,10] Such an approach, however, fails to account for potential confounding (*i.e.*, a variable having a spurious association with the outcome variable because it is also associated with an independent variable). A small number of studies have employed multivariate regression models—which can, in theory, address confounding—to identify associations between explanatory variables and levels of *E. coli* in stored water and on hands. For example, Levy *et al.* used a generalized estimating equations (GEE) model and found water source type, water treatment practices, and storage time to be significantly associated with levels of *E. coli* in stored water in Ecuadorian households.[11] Also using GEE models, a second study in Ecuador by Levy *et al.* (2009) found source type, water storage practices, and rainfall to be significantly correlated with *E. coli* in stored water.[12] Pickering *et al.* (2010) used linear regression and generalized estimating equations in their investigation of household characteristics and behaviors associated with levels of *E. coli* contamination of stored water and female caregiver hands in Tanzanian households.[5] Contamination on female caregiver hands was the only independent variable found to be statistically significantly associated with stored water quality. Having an infant present in the household was associated with higher hand contamination and educational attainment of mother was associated with lower hand contamination.

Substantive conclusions regarding the correlates of *E. coli* contamination in stored drinking water thus vary across these investigations. For example, water treatment was statistically significantly associated with stored water quality in just one of the studies described above; water source type was significant in two of the studies. Several plausible and non-mutually exclusive explanations exist for such divergence. The differences could reflect variation in the true relationship between water quality and independent variables across the study sites. In addition, authors did not evaluate the same independent variables in all analyses. It is also possible that limited variation in the values of independent variables caused them to be omitted or found not to be statistically significant in some analyses.

Whereas the substantive conclusions across these studies were compared, goodness-of-fit measures were not because comparable measures were not reported for the majority of the models. Only the hand contamination linear regression model in Pickering *et al.* (2010) reported an indicator reflecting model fit.[5] The model was characterized by poor explanatory power, with only 3% of the variation in the level of *E. coli* on two hands explained by the independent variables included in the model.

Researchers who employ linear regression techniques assume that the underlying relationship between log-transformed fecal indicator bacteria concentration and each explanatory variable is best represented by a straight-line function. If the effect of an explanatory variable is believed to be moderated by another variable, then interaction effects must be modeled explicitly, which has implications for sample size requirements. It may be, however, that the relationship between fecal indicator bacteria contamination in water and on hands and commonly tested correlates (*e.g.*, water management and hygiene behaviors, types of water sources and sanitation facilities used) is better characterized by nonlinear functions. For example, continuous explanatory variables and outcome measures may exhibit threshold effects. In such cases, the two variables are associated only above or below certain quantitative limits, or thresholds. Alternatively, such relationships could be characterized by equifinality (multiple causal pathways resulting in the same level of contamination), in which different combinations of explanatory variable values would be associated with the same level of contamination. Exploring other analytical methods that relate explanatory variables to water and hand contamination in fundamentally different ways could offer insight into these complex relationships.

Logistic regression and classification tree analysis are two such alternative analytical approaches. Logistic regression relates independent variables to the predicted probability of a categorical outcome, which can be binary or have three or more ordered or non-ordered values. Logistic regression has been used to model categorical outcomes in many different fields; for example, it has been applied extensively in medicine to model the probability of illness as a function of disease risk factors.[13] Each parameter estimate produced by logistic regression modeling can be interpreted as the average effect of a unit change in a correlate value on the predicted probability of a case belonging to a particular outcome category. Applied to fecal indicator bacteria contamination, logistic regression thus includes an assumption of threshold effects between contamination and an independent variable. At the same time, logistic regression methods still assume that the log-transformed predicted probabilities for a given outcome category are linearly related to each independent variable.[14] This modeling technique also assumes that the explanatory variables are independently related to the outcome, and tests for multicollinearity must be conducted in order to ensure this assumption holds for the data analyzed.

Classification tree analysis is a non-linear method that assigns cases into outcome categories based on the values of the explanatory variables, which can be continuous or categorical.[15] Classification tree models are particularly useful for identifying non-linear relationships and interactions among explanatory variables, as well as for predicting outlier cases.[15–17] Classification trees have been used to predict recreational water quality;[16,18–20] they have also been applied, either for prediction or identification of correlates, in medicine, pharmacology, ecology, computational biology, and bioinformatics.[15,17,21] A recent investigation employed classification tree analysis to predict soil-transmitted helminth infections using water, sanitation, and hygiene indicators.[22] The study identified latrine structure and cleanliness as the only predictors of infection.[22]

In this study, we employ logistic regression and classification trees, along with ordinary least squares regression, to explain

variation of fecal indicator bacteria concentration in stored drinking water and hand rinse samples. We use data collected from 1217 female heads of household in Bagamoyo, Tanzania, to estimate these three models. The primary objective of the study is to compare and contrast the explanatory variables selected in each of these analysis methods. We also assess the performance of the different models in terms of explaining variation in the outcome measures, and discuss the practical implications of our findings. These models are not optimized for prediction, and thus should not be used to predict outcomes for new sample data.

2. Methods

2.1. Study area and household selection

Households ($n = 1217$), each with at least one child under the age of five years and a female caretaker, were recruited for the study using cluster-randomized sampling from 15 different villages within the Bagamoyo District of Tanzania, East Africa (6°28′S 38°55′E). Data for this study were collected between March and May of 2010, during the baseline phase of a larger behavioral intervention study conducted in the area. Household visits consisted of an in-person interview with the female caretaker, as well as the collection of stored drinking water and hand rinse samples from each household. The research was approved by the Stanford Human Subjects Research and IRB (CA, USA) and the National Institute for Medical Review (NIMR) of Tanzania (Dar es Salaam, TZ). Free and informed consent was obtained for each participating household in the study. A subset of the households ($n = 93$) were included in a previously published study on hands and water as vectors for diarrheal pathogens.[7]

2.2. Household interview

Trained local enumerators conducted interviews with the primary female caregiver of each household. Information about socio-demographic characteristics along with water, sanitation, and hygiene behaviors and household health information were collected from each household and recorded on a handheld computer using a questionnaire developed with The Survey System (TSS) (Creative Research Systems, Petaluma, CA).

2.3. Water sample collection

Stored drinking water samples were collected from the household, with the respondent extracting the water from the storage container as she normally would and placing it in a sterile sampling bag (VWR, Radnor, PA). The sample was tested for chlorine using strips (Hach Company, Loveland, CO); sodium thiosulfate was added to any sample that tested positive for chlorine, in order to prevent chlorine-induced *E. coli* cell inactivation. Information about the stored water sampled was collected from the respondent, including the source from which it was obtained; length of time in storage; and whether it had undergone treatment. The enumerator also noted each respondent's method of extracting water from the storage container. All water samples were sealed and placed in a cooler on ice, then transported to the laboratory for microbial analysis within 6 h of collection.

2.4. Hand rinse sample collection

A hand rinse sample from the female caretaker was collected following previously published methods.[5] The respondent was asked to place her hands, one at a time, in a sterile sampling bag filled with 350 mL of distilled water. Information regarding the respondent's hand hygiene behaviors was collected, including time since last hand washing with soap, activity prior to hand rinse, and how the respondent typically wets and dries her hands for hand washing. One field blank of the hand rinse sampling bags (*i.e.*, taken to the field in the sample coolers, and then handled just like the samples) was processed each week to ensure that no contamination occurred during sample transport. All hand rinse samples were sealed and placed in a cooler on ice, then transported to the laboratory for microbial analysis within 6 h of collection.

2.5. Sample processing

In the water microbiology lab at the Ifakara Health Institute in Bagamoyo, all of the water and hand rinse samples were processed for the detection and enumeration of *E. coli* using membrane filtration, following USEPA Method 1604.[23] Stored water volumes of 100 mL were processed using membrane filtration, unless the turbidity of the sample prevented filtering the entire volume (3% of samples). Volumes of 1 mL and 10 mL were filtered for all hand rinse samples. If *E. coli* was not detected in a sample (*i.e.*, no colony forming units, or CFU, visible on filter after incubation), then half of the detection limit (1 CFU per plate) was used to calculate the concentration of *E. coli* per 100 mL or 2 hands rinsed. If a filter was too numerous to count (>500 CFU per filter) then 500 CFU was used to calculate the concentration of *E. coli* per 100 mL or 2 hands rinsed. Four or five method blanks were processed each day, and 10% of the samples were processed in duplicate.

2.6. Statistical analysis

Three modeling techniques were used to explain variation in stored water quality of study households: ordinary least squares regression, multinomial logistic regression, and classification tree. The outcome variable of the ordinary least squares regression model was log-transformed *E. coli* concentration, reported as colony-forming units (CFU) per 100 mL. For the categorical outcome indicator used in the multinomial logistic regression and classification tree models, water quality was classified by categories of "low" contamination (0–10 CFU *E. coli* per 100 mL), "medium" contamination (11–100 CFU *E. coli* per 100 mL), and "high" contamination (greater than 100 CFU *E. coli* per 100 mL). These categories were selected because they have been associated with both levels of health risk and water treatment guidelines for emergency situations.[24–26] A complete list of explanatory variables included in the models are displayed in Table 1. These variables were selected because they have the potential to impact the spread of fecal contamination (*e.g.*, water, sanitation, and hygiene related variables) or they are

Table 1 Study sample characteristics

Household characteristics	
Number of households in study	1217
Median weekly expenditures per capita for household, Tsh (USD)	6500 (4.3)
Female head of household works outside of home	22%
Female caregiver completed primary education	73%
Child 1 year old or less present in household	29%
GI illness[a] in household in past 48 h	9%
Household has dirt floor in home	51%
Household located within Bagamoyo town	61%
Household sanitation	
Household has private latrine	61%
Household latrine has a roof	27%
Household latrine has cement floor	35%
Household latrine has a septic tank	7%
Household latrine has a pit cover	21%
Feces visible around household premises	6%
Children in household practice open defecation	58%
Five or more flies visible in latrine	25%
Household water	
Household has JMP improved water source	82%
Water source on household premises	14%
Household actively treated drinking water	15%
Water sampled was stored less than 24 h	24%
Water storage container fully covered at time of sampling	95%
Drinking water extraction method risky	89%
Respondent hand contacted water during stored water extraction	15%
Household hand hygiene	
Household has hand washing station with soap and water	19%
Female caregiver dries hands with fabric after handwashing	65%
Female caregiver pours water from jerrycan to wet hands for handwashing	15%
Activity prior to hand rinse sample – washing	14%
Activity prior to hand rinse sample – food handling	20%
Activity prior to hand rinse sample – sitting	60%
Activity prior to hand rinse sample – other	6%
Median time since last hand washing with soap, hours	3
Household microbial indicators	
Geometric mean CFU EC per 100 mL in stored water	33
Percent households with 0–10 CFU EC per 100 mL in stored water	28%
Percent households with 11–100 CFU EC per 100 mL in stored water	35%
Percent households with more than 100 CFU EC per 100 mL in stored water	37%
Geometric mean CFU EC per 2 hands of female caregiver	263
Percent household with EC detected on female caregiver hands	72%

[a] GI illness described as 3 or more loose and watery stools in a 24 h period.

potential confounders (*e.g.*, weekly expenditure, education) as identified in prior work.[5] For predictive comparisons between models, the predicted log-transformed concentrations from the ordinary least squares regression model were classified into the water quality categories that match the categorical outcomes of the other models.

The same three techniques were also used to analyze female caregiver hand contamination. Concentrations of *E. coli* in the hand rinse sample were reported per 2 hands rinsed and log-transformed for use as the dependent variable in the ordinary least squares regression model. The outcome measure of hand contamination used in the logistic regression and classification tree models was whether *E. coli* was detected (1) or not (0) in the rinse sample. Because no risk thresholds for hand contamination were found in the literature, this binary category of *E. coli* detection was employed. Each predicted concentration of hand contamination generated by the ordinary least squares regression model was also classified as "detect" or "non-detect" of *E. coli* based on the lower detection limit of the hand rinse assay (17.5 CFU per 2 hands) to match the categorical outcome of the logistic regression and classification tree models.

Ordinary least squares regression and logistic regression modeling were conducted using PASW Statistics (SPSS Inc., Chicago, IL). These modeling techniques assume explanatory variables are independently related to the outcome. To ensure variables in the models did not violate this assumption, tests for multicollinearity were performed by reviewing tolerance and variance inflation factors (VIFs) of explanatory variables and correlations between explanatory variables. If tolerances were above 0.2, VIFs were less than two, and Pearson's r correlations were less than 0.8 between variables, then the model was assumed not to be compromised by multicollinearity.[14] Neither model was found to be compromised by multicollinearity. For the ordinary least squares regression models, we also tested that residuals were normally distributed with a predicted probability (*P–P*) plot and found no concerning deviations from the normality line. Explanatory variables for the models were chosen *a priori* based on theory and prior published research. For the regression models, all *a priori* chosen explanatory variables were included in the first run of the model. A reduced model was then estimated by an iterative process, first removing variables with $p > 0.20$, and then keeping only variables with $p < 0.10$. The results of the full models (*i.e.*, including all the *a priori* explanatory variables) are found in the ESI.†

Within-sample predictive power was estimated for the models as a secondary measure of comparison of model performance (see ESI†), since explanatory power (as measured by the coefficient of determination, R^2) could not be evaluated for all three modeling techniques. If developing a model for predictive purposes, predictive power should be evaluated on a new 'test' sample set (*i.e.*, not the sample set that built the model), as within-sample predictive power could overestimate performance.[27] However, the main objective of this study was to develop explanatory models, so for model development, decisions were made to optimize the statistical power (*i.e.*, large sample size prioritized over withholding data for a 'test' set).

Classification tree modeling was conducted using MATLAB & Simulink R2010a, version 7.10 (The MathWorks Inc., Natick, MA) using the 'classregtree' command. The classification tree starts with a parent-node containing all observations; the tree

is then split into two child-nodes based on an explanatory variable and the corresponding binary decision (*i.e.*, yes/no or threshold), such that cases in the two child-nodes are the most homogeneous with respect to the outcome variable, *i.e.*, the error cost (percent of cases incorrectly classified) is minimized. The child-nodes then become parent-nodes to undergo further splitting, and the process is repeated until a chosen tree optimization parameter is met. Each node in the tree is assigned a pruning level (*i.e.*, a level of branching) based on its associated error cost; nodes closer to the top of the tree have a higher pruning level. Therefore, the explanatory variables selected at higher pruning levels sort more cases into the correct outcome categories (*i.e.*, have a lower error cost).

In this study, classification trees were optimized by setting the 'minleaf' parameter, which is the minimum number of cases in a child-node required for branching of the tree to continue. The optimal 'minleaf' value was determined by minimizing the pooled error cost of a 10-fold cross-validation of the tree. A 10-fold cross-validation of the tree divides the full dataset into 10 equal sets, uses 9 of the 10 sets to train a model, and then calculates an error cost with the set that was not used to train the model (*i.e.*, test set). The error cost calculation is then repeated, alternating the set used as the test set, and a pooled error cost is calculated. Values of 'minleaf' parameters from 1–100 were tested to find the value associated with the minimum pooled error cost of a 10-fold cross-validation of the tree.

3. Results

3.1. Blanks and duplicates

There were a total of 230 method blanks processed during the study. There were two blanks with contamination detected on the MI plate (*E. coli*-specific media). One CFU grew on the two contaminated MI plates, which were from blanks processed on different days. All other blanks processed on those days had zero growth (*i.e.*, no contamination). Eight field hand bag blanks were processed and found to be negative for contamination. Also, ten hand bag lab blanks (from hand bags that did not go into the field) were processed and found to be negative for contamination. As contamination was at very low concentrations and occurred in <1% of blanks, we did not correct for the contamination and included all data collected in the analysis. For the majority of samples processed in duplicate (>95%), the duplicate samples had fecal indicator bacteria concentrations of the same order of magnitude. The first sample processed of the duplicate pair was used in the analysis.

3.2. Sample household characteristics

Table 1 shows descriptive data of the 1217 sample households. The average household size was 5.5 people, and households had a median weekly per capita expenditure of 4.33 USD. Eighteen percent of households reported having access to an improved sanitation facility as defined by the Joint Monitoring Program.[28] The majority of households reported using an improved water source (82%), including piped water (60%), borewells (16%), and rainwater (5%). Only 13% of households had access to an improved water source located on their premises (*i.e.*, on-plot). The majority of water samples collected from the households were contaminated with fecal indicator bacteria, with 28% having 0–10 CFU *E. coli* per 100 mL, 35% having 11–100 CFU *E. coli* per 100 mL and 37% having greater than 100 CFU *E. coli* per 100 mL. The relationship between source water type and stored water quality is further detailed in Table 2. The majority of stored water samples (68%) originating from improved water sources (whether on-plot or off-plot) had medium or high *E. coli* levels. Notably, whether the improved water source was on-plot or off-plot, the stored water samples had low *E. coli* levels for the same fraction of households (32%) (Table 2). In addition, 15% of respondents reported treating the sampled stored drinking water, either by boiling, filtering, or chlorination. More than half (52%) of these households had highly contaminated stored drinking water. Almost three quarters of female caregivers were found to have *E. coli* in their hand rinse samples.

3.3. Stored water quality: identified correlates

The results of the ordinary least squares regression, multinomial logistic regression, and classification tree stored water models are presented in Table 3, and a diagram of the classification tree model is shown in Fig. 1. Only two variables were identified as explanatory variables of *E. coli* concentrations in all three models. First, having stored water that was obtained from an improved *versus* unimproved source was strongly associated with higher stored water quality. For instance, stored water source classification had three times the average effect on *E. coli* concentrations as compared to water storage time classification (more or less than 24 h). Second, the concentration of *E. coli* on female caregiver's hands was also identified as an explanatory variable of stored water contamination in all 3 analyses. In addition, several variables were found not to be explanatory variables of stored water *E. coli* concentrations in any model: whether the household's water source was located on the premises, whether the water storage container was covered at the time of sampling, and whether the respondent extracted

Table 2 Percent (and number) of households in each stored water quality category by JMP classified access to improved water source

	Stored water contamination categories		
	Low, 0–10 CFU EC per 100 mL	Medium, 11–100 CFU EC per 100 mL	High, >100 CFU EC per 100 mL
Unimproved water source	10.6%(23)	33.6%(73)	55.8%(121)
Improved water source off-plot	32.1%(264)	33.7%(277)	34.3%(282)
Improved water source on-plot	32.1%(51)	40.3%(64)	27.7%(44)

Table 3 Comparison of reduced models explaining stored water quality of households

	Ordinary least squares regression[a]		Multinomial logistic regression: medium EC category[b]		Multinomial logistic regression: high EC category[b]		Classification tree[c]
Variable[d]	B[e]	SE	B	SE	B	SE	Prune level
Constant	1.8	0.1	0.52	0.40	1.16***	0.37	—
Respondent works outside the home[f]	−0.20***	0.07	−0.45**	0.18	−0.50***	0.18	—
Regular weekly expenditure per capita	—	—	—	—	—	—	4
House has dirt floor[f]	0.16***	0.06	—	—	—	—	—
House located within town[f]	−0.13**	0.06	−0.30*	0.17	−0.41**	0.17	—
Infant present in household[f]	—	—	0.21	0.18	0.40**	0.17	—
Household has private latrine[f]	—	—	0.41**	0.16	0.21	0.16	—
Feces visible around household[f]	—	—	0.29	0.36	0.59*	0.35	—
Latrine has a cement floor[f]	—	—	—	—	—	—	5
Children open defecate[f]	—	—	—	—	—	—	0
Water source is improved[f]	−0.52***	0.07	−0.95***	0.26	−1.56***	0.25	2
Water was actively treated[f]	—	—	—	—	—	—	2
Water extracted in risky manner[f]	—	—	0.46*	0.25	0.04	0.23	—
Hand contacted water when extracting[f]	0.14*	0.07	0.19	0.23	0.51**	0.22	—
Water stored for less than 24 h[f]	−0.16**	0.06	−0.40**	0.18	−0.25	0.18	—
Log EC CFU per 100 mL on hands of caregiver	0.08***	0.03	0.06	0.08	0.19**	0.08	3

[a] Dependent variable is log CFU EC per 100 mL water. [b] Reference group is low contamination level category. [c] Outcome categories are low, medium, and high EC contamination categories. [d] Variables tested, found not to be significant, and excluded from models include: someone in the household has GI illness, latrine has a septic tank, latrine has a roof, latrine has a pit cover, flies present in latrine, water source on-plot, and water storage container is covered. [e] Unstandardized beta coefficient. [f] Binary variable (0 or 1). ***$p < 0.01$ **$0.01 \geq p < 0.05$ *$0.05 \geq p < 0.10$.

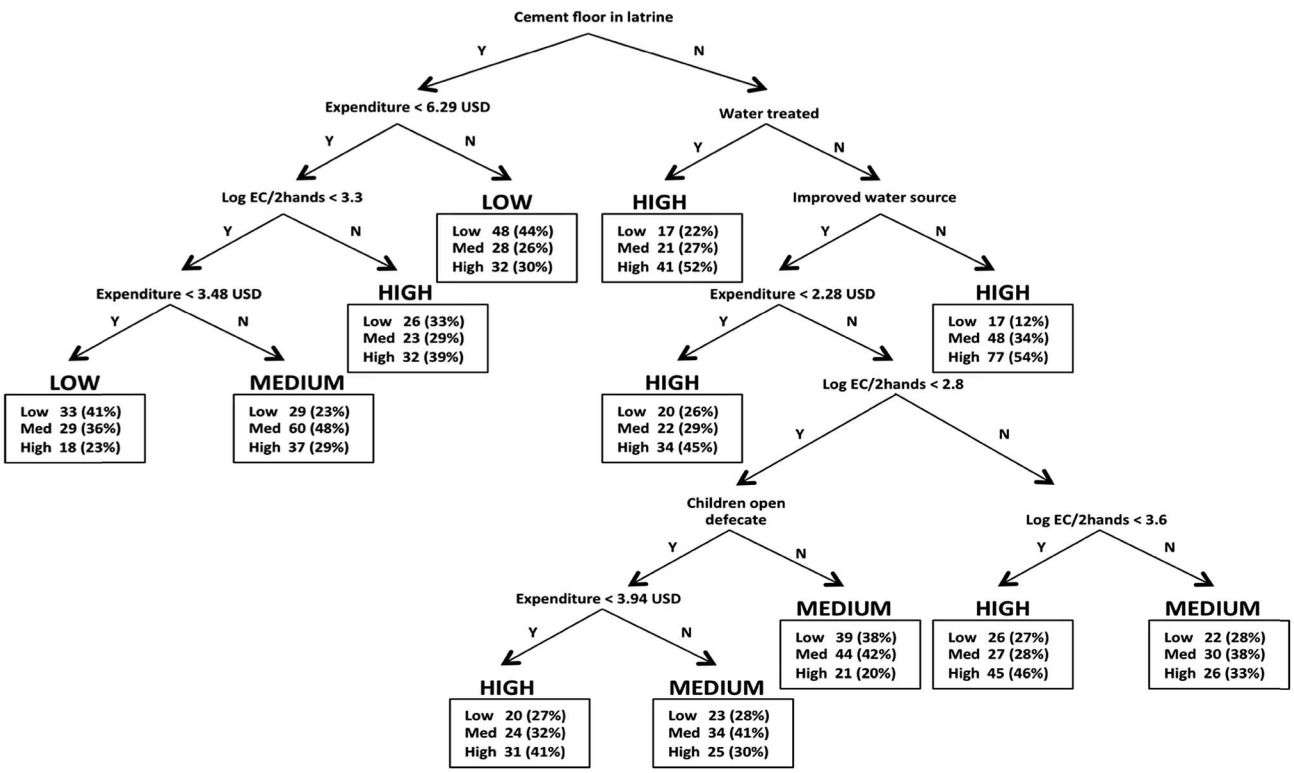

Fig. 1 Classification tree for low (0–10 CFU EC per 100 mL), medium (11–100 CFU EC per 100 mL), and high (greater than 100 CFU EC per 100 mL) levels of EC in stored drinking water sample. The predicted category of contamination is in bold above the terminal boxes. The terminal boxes report the distribution of cases in the terminal node. Minleaf = 75 cases. Expenditure is regular weekly expenditure per capita for the household in Tanzanian shillings (Tsh.). Log EC per 2 hands is the concentration of *E. coli* per two hands in female caregiver hand rinse samples. 'Water treated' refers to any form of active treatment (*i.e.*, boiling, chlorination, filtration, adding coagulant). Boxes reveal the number (and %) of cases in the terminal node by contamination category.

water from the storage container using a cup or bowl (*versus* a long-handled utensil or decanting).

For some associated correlates, however, the models did not agree. For instance, reported active treatment of the water, which included boiling, adding a coagulant, filtration, and chlorination, was not statistically significant in the ordinary least squares regression or multinomial logistic regression models but was a classification node in the classification tree model. Having a latrine with a cement floor was the first classification node in the classification tree model, meaning it was the variable that best sorts the cases into homogeneous groups of the outcome variable.[16] This variable was not statistically significant in either regression model, however. The modeling techniques relate explanatory variables to stored water quality in fundamentally different ways, as mentioned in the introduction, so it is not surprising that differences arise in terms of explanatory variables identified. However, based on underlying theory, we are not able to determine which of the models is 'correct' (*i.e.*, reflecting the true state of the world) but rather, each model, can offer insight for future hypothesis generation.

Uniquely, the classification tree model identified several 'recipes'—or combinations of characteristics—for households with highly contaminated water (Fig. 1). For instance, the predicted probability of high contamination in stored water was 0.54 for a household without a cement floor in the latrine, that did not report treating their water, and that obtained their water from an unimproved source. Interestingly, if a household did not have a cement floor but reported that they did treat their water, the model still predicted high contamination. The classification tree model predicted low contamination in the stored drinking water for households that had a cement floor in the latrine and regular weekly expenditure greater than 6.29 USD (47% of households reported greater than 6.29 USD regular weekly expenditure). Also, the model predicted low contamination for households that had a cement floor in the latrine, female caregiver hand contamination less than 3.3 log CFU *E. coli* per 2 hands, and weekly expenditure of less than 3.48 USD (39% of households reported less than 3.48 USD regular weekly expenditure).

3.4. Stored water quality: comparison of analytical techniques

The ordinary least squares regression model had a multiple R^2 value of 0.09, and the multinomial logistic regression had a Cox-Snell pseudo-R^2 of 0.10. In both cases, having values close to zero indicates poor model fit; however, the parameters are calculated differently and do not have the same mathematical interpretation.[14] An analogous parameter cannot be computed for classification trees. To allow for some comparison of model performance with the classification tree, within-sample predictive power was also estimated (see ESI, Table S1†).

3.5. Female caregiver hand contamination: identified correlates

The results of the three models for female caregiver hand contamination are shown in Table 4; Fig. 2 shows a diagram of the classification tree model. The only explanatory variable in all three models was the type of activity in which the respondent was engaged just prior to her hand rinse sample being taken (*e.g.*, washing, handling food, sitting, or 'other' activity). Preparing food and washing dishes, clothes, hands or children were positively associated with female caregiver hand contamination.

Table 4 Comparison of reduced models explaining detection of *E. coli* on female caregiver hands

	Ordinary least squares regression[a]		Binary logistic regression[b]		Classification tree[c]
Variable[d]	B[e]	SE	B	SE	Prune Level
Constant	2.30***	0.09	0.48**	0.23	—
Respondent works outside the home[f]	—	—	0.36**	0.17	—
Regular weekly expenditure per capita[g]	−0.02***	0.01	—	—	1
House located in town[f]	0.39***	0.06	0.58***	0.14	1
Infant present in household[f]	0.16**	0.07	—	—	—
Household has private latrine[f]	−0.16**	0.06	−0.33**	0.14	—
Feces visible around household[f]	0.25*	0.13	0.84**	0.34	—
Latrine has a cement floor[f]	—	—	—	—	2
Latrine has a septic tank[f]	−0.28**	0.12	—	—	—
Flies present in latrine[f]	—	—	−0.26*	0.15	—
Children open defecate[f]	—	—	0.27**	0.14	—
Time since last hand washing 1 h or less[f]	—	—	—	—	3
Prior activity involved washing[h]	0.19**	0.09	0.47**	0.21	—
Prior activity food handling[h]	—	—	0.36**	0.18	—
Prior activity (for classification tree only)[i]	—	—	—	—	1

[a] Dependent variable is log CFU *E. coli* per 2 hands. [b] Reference group is no detection of *E. coli*. [c] Outcome categories are *E. coli* detected or not on female caregiver hands; pruning level represents the level of branching in the tree with nodes at the top of the tree having a higher pruning level. [d] Variables tested, found not to be significant, and excluded from models include: house has a dirt floor, someone in household has GI illness, latrine has a roof, latrine has pit cover, respondent has primary education, hand washing station with soap present, hands dried with fabric after hand washing, and hands wetted for hand washing by pouring water. [e] Unstandardized beta coefficient. [f] Binary variable (0 or 1). [g] In (1000 Tsh). [h] Dummy variables with the reference activity of 'sitting' and 'other activities'. [i] Categorical variable of activity prior to hand rinse being sitting, washing, food handling, or other. ***$p < 0.01$ **$0.01 \geq p < 0.05$ *$0.05 \geq p < 0.10$.

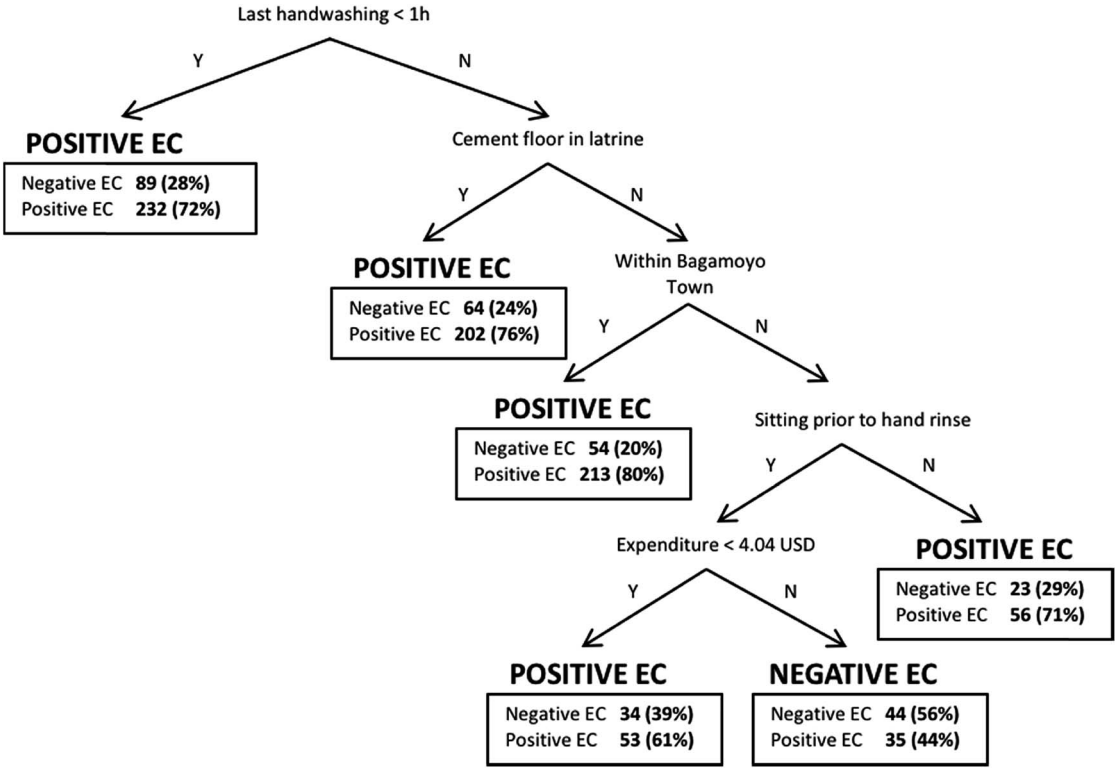

Fig. 2 Classification tree for *E. coli* (EC) detection on the hands of female caregivers (*i.e.,* positive or negative for detection of EC). Female caregiver hand rinses detecting EC (*i.e.,* positive EC) and not detecting EC (*i.e.,* negative EC). The predicted category of contamination is in bold above the terminal boxes. The terminal boxes report the distribution of cases in the terminal node. Minleaf = 75 cases. Last hand washing is time since last hand washing with soap. 'Within Bagamoyo Town' is whether or not the household is located within Bagamoyo town. Activity prior to hand rinse is sitting as opposed to any other activity. Expenditure is regular weekly expenditure per capita for the household in US dollars (USD). Boxes reveal the number (and %) of cases in the terminal node by contamination category.

Other variables were only identified as correlates in one or two of the models. A household having a private latrine, as well as feces being observed on the ground near the household, were both statistically significantly associated with higher levels of *E. coli* contamination on a respondent's hands in the ordinary least squares regression and binary logistic regression models. For the ordinary least squares regression model, the household having a septic tank was associated with decreased contamination on the respondent's hands. For the binary logistic regression model, household children reportedly practicing open defecation and feces being visible around the household were statistically significantly associated with increased contamination on the respondent's hands. Time since last hand washing with soap was the first node in the classification tree (*i.e.*, highest pruning level), meaning it was the most effective variable in sorting cases into homogenous groups of *E. coli* detect and *E. coli* non-detect (Fig. 2). Interestingly, the tree predicts *E. coli* detection in the hand rinse sample if the respondent reported hand washing with soap less than 1 hour prior to sampling. The classification tree model only had one branch that predicted no detection of *E. coli* in the rinse sample, and it was for respondents that reported hand washing more than 1 hour prior to rinse, do not have a cement slab on their latrine, live outside of Bagamoyo town, were sitting prior to the rinse sample, and had comparatively high regular weekly expenditures.

3.6. Female caregiver hand contamination: comparison of analytical techniques

For traditional fit indexes, the ordinary least squares regression model has a multiple R^2 value of 0.06, and the binary logistic regression has a Cox-Snell pseudo-R^2 of 0.05 (an analogous statistic cannot be calculated for classification trees). The within-sample predictive power is shown in the ESI (Table S3†).

4. Discussion

The three analytical methods used to describe microbial contamination on female caregiver hands and in stored drinking water of households make fundamentally different assumptions about the relationships between explanatory variables and outcome measures of contamination. In particular, the classification tree identified complex clusters of explanatory variables associated with a given outcome. By contrast, the regression models generate estimated average effects for each explanatory variable. As a result, despite using the same data set, different explanatory variables were identified across the models tested. Thus, substantively different

conclusions were drawn based on the model selected. As analysis of cross-sectional data is useful for the generation of hypotheses, the model selected may influence which hypotheses are recommended for future testing. Selection of the appropriate analytical technique requires understanding the relationship between explanatory and outcome variables. For example, forcing a linear relationship between explanatory and outcome variables when a threshold relationship exists would increase the standard error and represent an inaccurate magnitude of effect associated with the explanatory variable. Also, models have different assumptions, such as independence of co-variates in regression models, that must not be violated. To appropriately guide future research, the model selection should reflect the understanding of how the explanatory variables are believed to be related to the outcome. If the relationships between explanatory variables and the outcome are unclear, such as the case with water and hand contamination, then using multiple modeling techniques can be fruitful. However, it is worth noting that classification tree does allow for more complex relationships to be identified between explanatory variables, and this method exhibits improved within-sample predictive power, particularly for outlier cases. Additional modeling techniques, such as neural network analysis and support vector machines, could be explored to explain variation in stored drinking water and hand contamination.

There were some explanatory variables that were identified as related to the outcome variables across the three modeling types (*e.g.*, water source type for the water models and activity prior to sampling for the hand models), providing triangulated support for the relationships. Strikingly, whether an improved water source is on-plot or off-plot doesn't result in improved water quality outcomes (Table 2). As an on-plot improved water source represents the top of the water ladder,[28] this study highlights that if drinking water is still stored in the home, the achievement of access to improved water infrastructure on the living premises would not necessarily confer water quality gains.

The explanatory power of the regression models was lower for hand contamination than for stored water quality, and was on par with previous research.[5] In addition, all three modeling techniques exhibited low *within-sample* predictive power (see ESI†). Similar to other studies,[5,11,12] our results highlight the complexity of explaining stored water quality and hand contamination in low-resource settings.

Several possible explanations exist for the poor explanatory and within-sample predictive power of these models. The limited variation in values of the outcome variable (*e.g.*, low percentage of cases were non-detects for the hand contamination categorical outcome models) could contribute to the poor model performance.[22] Also, several of the explanatory variables used in the analyses were based on self-reported data. Biased responses could prevent the identification of an existing relationship between the outcome variable and a correlate. In particular, unreliable reporting of hygiene behaviors has been documented in other studies.[29,30] Aside from biases in the collected data, omission of important correlates could limit the predictive power of the models.

Poor model performance could also stem from *E. coli* concentrations not being an appropriate indicator of fecal contamination in drinking water and on hands. For instance, *E. coli* have been found to be naturally occurring in soils (*i.e.*, not from feces) in tropical environments.[31–33] In such a case, one would not necessarily expect the concentration of these organisms in stored water to be correlated with household water management and hygiene practices. Additionally, *E. coli* are found in the feces from multiple animal hosts, not just humans, which is problematic since many of the sanitation-related variables (*e.g.*, children practicing open defecation) included in the models focus on contamination from human, rather than non-human feces.

We also note that *E. coli* measurements can exhibit considerable intrinsic sampling variability.[34] Such random sampling error can impede efforts to identify associations between *E. coli* and extrinsic explanatory variables using multivariate statistical modeling.[35] In theory, taking replicate water quality measurements would reduce measurement error, improve model fit, and increase precision of parameter estimates.[35] Future research that explores the impact of replicate sampling on the explanatory power and parameter estimates of water quality and hand contamination models would thus be a valuable contribution. Some researchers have applied the Spearman–Brown formula to determine the number of replicate measures needed for a desired level of precision in a parameter estimate.[36]

Variability in *E. coli* measurements within a household may also be non-random. For example, there is evidence that water quality varies systematically over both short (<1 day) and longer (>1 day) time scales.[12,37] The water and hand samples taken in the present study were captured at different times of day, and no information was available regarding temporal trends in contamination among study households. In future research, incorporating both repeat measurements within a household (over a relevant time frame) and replicate measurements (at each point in time) could allow modeling efforts to estimate the share of total variance attributable to explanatory variables, systematic temporal variation, and random variability.

5. Conclusions

Tests of microbial contamination in household environments has been expanding, particularly in low-income country settings, in effort to understand disease transmission pathways. However, as evidenced in this work, analytical techniques used to identify covariates or risk factors for contamination outcomes were limited in their capacity to explain variation in stored drinking water quality and female caregiver hand contamination. Measurement error and non-random temporal variability that is not captured by parameter estimates may be contributing to the unexplained variation. Future research should explore how different water sample schemes (such as repeated measures at different temporal scales) can improve explanatory power of models.

Despite some limitations in the models, this study did provide some insight between behaviors and household characteristics and contamination in water and on hands that could

be further explored for causal relationships in experimental evaluation. This work highlighted how the use of multiple modeling techniques can be fruitful when underlying relationships between explanatory variables and an outcome remain unclear. Although the modeling approaches varied, there were some similarities in the results, with certain explanatory variables being consistently identified as being related to hand and water contamination (*e.g.*, water source type for the water models and activity prior to sampling for the hand models). At the same time, there were also marked differences across the models. In sum, these results suggest there are benefits to using multiple analysis methods to assess relationships in complex systems.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This study was funded by the National Science Foundation (SES-0827384). Angela R. Harris was funded by the National Science Foundation Graduate Research Fellowship and the Stanford Graduate Fellowship while conducting this work. The authors would like to thank field staff and laboratory technicians at the Ifakara Health Institute for their support as well as participating households. The authors also acknowledge the support of Omar Juma, Maggie Montgomery, Emily Viau, Mia Mattioli, and Michael Harris.

References

1 Joint Monitoring Program, *Progress on Drinking Water, Sanitation and Hygiene*, World Health Organization and United Nations Children's Fund, Geneva, 2017, available from: https://www.unicef.org/media/media_96632.html.
2 J. Wright, S. Gundry and R. Conroy, Household drinking water in developing countries: a systematic review of microbiological contamination between source and point-of-use, *Trop. Med. Int. Health*, 2004, **9**(1), 106–117.
3 J. V. Pinfold, Faecal contamination of water and fingertip-rinses as a method for evaluating the effect of low-cost water supply and sanitation activities on faeco-oral disease transmission. I. A case study in rural north-east Thailand, *Epidemiol. Infect.*, 1990, **105**(2), 363–375.
4 A. F. Trevett, R. C. Carter and S. F. Tyrrel, Mechanisms leading to post - supply water quality deterioration in rural Honduran communities, *Int. J. Hyg. Environ. Health*, 2005, **208**(3), 153–161.
5 A. J. Pickering, J. Davis, S. P. Walters, H. M. Horak, D. P. Keymer, D. Mushi, *et al.*, Hands, water, and health: fecal contamination in Tanzanian communities with improved, non-networked water supplies, *Environ. Sci. Technol.*, 2010, **44**(9), 3267–3272.
6 L. Roberts, Y. Chartier, O. Chartier, G. Malenga, M. Toole and H. Rodka, Keeping clean water clean in a Malawi refugee camp: a randomized intervention trial, *Bull. W. H. O.*, 2001, **79**(4), 280–287.
7 M. C. Mattioli, A. J. Pickering, R. J. Gilsdorf, J. Davis and A. B. Boehm, Hands and Water as Vectors of Diarrheal Pathogens in Bagamoyo, Tanzania, *Environ. Sci. Technol.*, 2013, **47**, 355–363.
8 V. Curtis and S. Cairncross, Reviews effect of washing hands with soap on diarrhoea risk in the community : a systematic review, *Lancet*, 2003, **3**, 275–281.
9 S. Shrestha, S. S. Malla, Y. Aihara, N. Kondo and K. Nishida, Water Quality at Supply Source and Point of Use in the Kathmandu Valley, *J. Water Environ. Nanotechnol.*, 2013, **11**(4), 331–340.
10 a. Shaheed, J. Orgill, C. Ratana, M. a. Montgomery, M. a. Jeuland and J. Brown, Water quality risks of "improved" water sources: evidence from Cambodia, *Trop. Med. Int. Health*, 2014, **19**(2), 186–194.
11 K. Levy, K. L. Nelson, A. Hubbard and J. N. S. Eisenberg, Following the water: a controlled study of drinking water storage in northern coastal Ecuador, *Environ. Health Perspect.*, 2008, **116**(11), 1533–1540.
12 K. Levy, A. E. Hubbard, K. L. Nelson and J. N. S. Eisenberg, Drivers of water quality variability in northern coastal Ecuador, *Environ. Sci. Technol.*, 2009, **43**(6), 1788–1797.
13 S. C. Lemon, J. Roy, M. A. Clark, P. D. Friedmann and W. Rakowski, Classification and Regression Tree Analysis in Public Health : Methodological Review and Comparison with Logistic Regression, *Ann. Behav. Med.*, 2003, **26**(3), 172–181.
14 A. Field, *Discovering statistics using IBM SPSS Statistics*, Sage Publications Inc., London, 4th edn, 2013.
15 G. De'ath and K. Fabricius, Classification and regression trees: a powerful yet simple technique for ecological data analysis, *Ecology*, 2000, **81**(11), 3178–3192.
16 H.-K. Bae, B. H. Olson, K.-L. Hsu and S. Sorooshian, Classification and regression tree (CART) analysis for indicator bacterial concentration prediction for a Californian coastal area, *Water Sci. Technol.*, 2010, **61**(2), 545–553.
17 C. Strobl, J. Malley and G. Tutz, An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests, *Psychol. Methods*, 2009, **14**(4), 323–348.
18 R. T. Stidson, C. A. Gray and C. D. McPhail, Development and use of modelling techniques for real-time bathing water quality predictions, *Water Environ. J.*, 2012, **26**(1), 7–18.
19 W. Thoe, M. Gold, A. Griesbach, M. Grimmer, M. L. Taggart and A. B. Boehm, Predicting water quality at Santa Monica Beach: evaluation of five different models for public notification of unsafe swimming conditions, *Water Res.*, 2014, **67**, 105–117.
20 W. Thoe, S. H. C. Wong, K. W. Choi and J. H. W. Lee, Daily prediction of marine beach water quality in Hong Kong, *J. Hydro-Environ. Res.*, 2012, **6**(3), 164–180.
21 C. Kingsford and S. Salzberg, What are decision trees?, *Nat. Biotechnol.*, 2008, **26**(9), 1011–1013.
22 K. Gass, D. G. Addiss and M. C. Freeman, Exploring the relationship between access to water, sanitation and

hygiene and soil-transmitted helminth infection: a demonstration of two recursive partitioning tools, *PLoS Neglected Trop. Dis.*, 2014, **8**(6), e2945.

23 USEPA. Method 1604, *Total Coliforms and Escherichia coli in water by membrane filtration using a simultaneous detection technique (MI medium)*, Washington, DC, 2002.

24 K. Onda, J. LoBuglio and J. Bartram, Global access to safe water: accounting for water quality and the resulting impact on MDG progress, *Int. J. Environ. Res. Public Health*, 2012, **9**(3), 880–894.

25 R. Baum, G. Kayser, C. Stauber and M. Sobsey, Assessing the microbial quality of improved drinking water sources: results from the Dominican Republic, *Am. J. Trop. Med. Hyg.*, 2014, **90**(1), 121–123.

26 *Environmental Health in Emergencies and Disasters: A practical guide*, ed. B. Wisner and J. Adams, World Health Organization, 2002, ch. 7 Water Supply, pp. 92–126.

27 G. Shmueli, To Explain or to Predict?, *Stat. Sci.*, 2010, **25**(3), 289–310.

28 Joint Monitoring Program, *Progress on Sanitation and Drinking Water: 2010 Update*, World Health Organization and the United Nations Children's Fund, Geneva, 2010.

29 V. Curtis, S. Cousens, T. Mertens, E. Traore, B. Kanki and I. Diallo, Structured observations of hygiene behaviours in Burkina Faso: validity, variability, and utility, *Bull. W. H. O.*, 1993, **71**(1), 23–32.

30 M. Manun'Ebo, S. Cousens, P. Haggerty, M. Kalengaie, A. Ashworth and B. Kirkwood, Measuring hygiene practices: a comparison of questionnaires with direct observations in rural Zaïre, *Trop. Med. Int. Health*, 1997, **2**(11), 1015–1021.

31 M. N. Byappanahalli, T. Yan, M. J. Hamilton, S. Ishii, R. S. Fujioka, R. L. Whitman, *et al.*, The population structure of Escherichia coli isolated from subtropical and temperate soils, *Sci. Total Environ.*, 2012, **417–418**, 273–279.

32 M. N. Byappanahalli, B. M. Roll and R. S. Fujioka, Evidence for Occurrence, Persistence, and Growth Potential of Escherichia coli and Enterococci in Hawaii's Soil Environments, *Microbes Environ.*, 2012, **27**(2), 164–170.

33 F. P. Brennan, J. Grant, C. H. Botting, V. O'Flaherty, K. G. Richards and F. Abram, Insights into the low-temperature adaptation and nutritional flexibility of a soil-persistent Escherichia coli, *FEMS Microbiol. Ecol.*, 2013, **84**(1), 75–85.

34 A. D. Gronewold and R. L. Wolpert, Modeling the relationship between most probable number (MPN) and colony-forming unit (CFU) estimates of fecal coliform concentration, *Water Res.*, 2008, **42**(13), 3327–3334.

35 J. M. Fleisher, The effects of measurement error on previously reported mathematical relationships between indicator organism density and swimming-associated illness: a quantitative estimate of the resulting bias, *Int. J. Epidemiol.*, 1990, **19**(4), 1100–1106.

36 J. A. Hutcheon, A. Chiolero and J. A. Hanley, Random measurement error and regression dilution bias, *BMJ*, 2010, **340**(7761), 1402–1406.

37 A. R. Harris, J. Davis and A. B. Boehm, Mechanisms of post-supply contamination of drinking water in Bagamoyo, Tanzania, *J. Water Health*, 2013, **11**(3), 543–554.