Supplemental Information for "Comparison of analytical techniques to explain variability in stored drinking water quality and microbial hand contamination of female caregivers in Tanzania"

Authors: Angela R. Harris*[1, 2], Amy J. Pickering[1, 3], Alexandria B. Boehm[1], Mwifadhi Mrisho[4], Jennifer Davis[1, 5]

Methods: Matlab Code for Classification Tree

Hand Model

```
clear
clc
count=0;

ML=75;

cat_var=[1 2 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20];

[data,text]=xlsread('Hand_FIBdata_FEB2019_Main.xlsx');
var_label=text(1,:);
var_labelc=char(var_label(1:20));
[row, col]=size(data);
independent=data(:,1:col-1);
dependent=data(:,col);
tree = classregtree(independent,dependent,'names', var_labelc,
'MinLeaf',ML,'categorical',cat_var, 'method', 'classification');

showTree = prune(tree,'level',0)
view(tree,'names',var_label);
```

Water Model

```
clear
clc
count=0;

ML=75;
cat_var=[1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21];

[data,text]=xlsread('SW_FIB_tree_Feb2019_Main.xlsx');
var_label=text(1,:);

[row, col]=size(data);
var_labelc=char(var_label(1:22));
independent=data(:,1:col-1);
dependent=data(:,col);
tree = classregtree(independent,dependent,'names',var_labelc,
'MinLeaf',ML,'categorical',cat_var, 'method', 'classification');


view(tree, 'names', var_label);
```

Methods: In-sample Predictive Power

To compare the performance of the different analytical techniques, the measures of the models' predictive capabilities (i.e., percent correctly predicted overall, sensitivity, and specificity) were compared using a test of proportions with an independent sample z-test. The percent of samples correctly predicted overall was calculated by dividing the number of cases correctly predicted by the total number of cases. The sensitivity of a model to predict a contamination category was calculated as

$$Sensitivity = \frac{True\ positives\ for\ category}{(True\ positives\ for\ category + False\ negatives\ for\ category)} \qquad (1)$$

where true positives are cases for which both the predicted value and the observed value are positive for a given contamination category. A false negative is a case for which the observed value is positive, but the model predicts negative, for a given contamination category. To determine if the stored water quality models predicted an outcome differently than would be expected by chance, the sensitivity was compared to 0.33 (e.g., the probability of selecting one category out of three at random—low, medium, or high). For the binary outcome of hand contamination, the specificity of a model was calculated as

$$Specificity = \frac{True\ negatives}{(True\ negatives + False\ positives)} \qquad (2)$$

where the true negatives are cases for which the model predicts negative for *E. coli* contamination on hands and the observed values are negative as well. A false positive is a case for which the model predicts positive for *E. coli* contamination on hands, but the observed value is negative. For all statistical tests, *p*-values less than 0.05 were considered statistically significant.

Results:

Table S1 presents information about the predictive capabilities of the three different stored water quality models. The ordinary least squares regression model correctly predicted household drinking water quality categories for 36% of cases overall, which is not statistically different from what would be expected by chance (p>0.05, difference in proportions independent sample z-test). Alternatively, the multinomial logistic regression predicted 44% of cases correctly, and the classification tree model predicted 46% of cases correctly; both models performed better than would be expected by chance (p<0.001, difference in proportions independent sample z-test). The ordinary least squares regression model correctly predicted the majority of cases with medium contamination (89%), and only predicted 1% and 14% of the low and high contamination categories correctly, respectively. The multinomial logistic regression model predicted the lowest contamination category better than the other models (45% of low contamination cases correctly predicted) and the classification tree model predicted the high contamination category better (62% of high contamination cases correctly predicted) (p<0.001, difference in proportions independent sample z-test).

Table S1 Comparison of model performance for predicting stored water contamination 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). Percent of cases in each category and in overall sample correctly predicted are reported. The ‘n’ number of cases included in each model is also reported, and this value varies by model due to the way each model technique handles missing values.

Model	**% low contamination correct**	**% medium contamination correct**	**% high contamination correct**	**% overall correct**
OLS (n=1185)	1%	88%	14%	36%
MLR (n=1143)	45%	34%	51%	44%
CT (n=1129)	25%	43%	62%	46%

The distribution of ‘low’, ‘medium’, and ‘high’ contamination for the different cases included in each of the models is shown in Table S2.

Table S2. Comparison of the number of observations with ‘low’, ‘medium’, and ‘high’ contamination in stored water for each model reported in Table S1.

Model	**Overall (N)**	**Low (N)**	**Medium (N)**	**High (N)**
MLR	1143	330	392	421
OLS	1185	335	408	442
CT	1129	320	390	419

As shown in Table S3, all three models predicting female caregiver hand contamination had sensitivities of 95% or above. By contrast, the absence of *E. coli* was poorly predicted by the models. The classification tree model correctly predicted 14% of the cases negative for *E. coli* in the hand rinse sample. The ordinary least squares regression model had a 0% specificity, because it never predicts concentrations below the limit of detection (i.e., non-detect of *E. coli*). For the hand rinse samples, the binary logistic regression model correctly predicted the classification of 3% of cases with *E. coli* not detected.

Table S3 Comparison of model performance for predicting detection of *E. coli* on female caregiver hands. Sensitivity is the fraction of cases positive for *E. coli* correctly predicted, and specificity is the fraction of cases with *E. coli* not detected correctly predicted. The ‘n’ number of cases included in each model is also reported, and this value varies by model due to the way each model technique handles missing values.

Model	**Sensitivity**	**Specificity**	**% Correctly Predicted Overall**	**True Positives**	**True Negatives**

OLS (n=1150)	100%	0%	74%	821	329
BLR (n=1163)	99%	3%	72%	834	329
CT (n=1099)	96%	14%	72%	791	308

In order to observe the influence of missing observations (sample data) on the performance of the different modeling techniques, all 3 models were re-run to include the same number of observations that had no missing data. In addition, the full variable list (i.e., not producing the reduced model that prioritizes explanatory power) was including in the models. The results are presented below (Table S4-S7).

Table S4 Comparison of model performance for predicting stored water contamination 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). Percent of cases in each category and in overall sample correctly predicted are reported. For this comparison, the models all include the same explanatory variables (i.e., no reduced form of the model) and the same number of observations (i.e., sample data used to construct the modes, ‘n’). The number of cases with observed ‘low’ contamination is 318, ‘medium’ is 382, and ‘high’ is 411.

Model	**% low contamination correct**	**% medium contamination correct**	**% high contamination correct**	**% overall correct**
OLS (n=1111)	4%	89%	16%	38%
MLR (n=1111)	41%	40%	55%	46%
CT (n=1111)	34%	32%	65%	45%

Table S5 Comparison of model performance for predicting detection of *E. coli* on female caregiver hands. Sensitivity is the fraction of cases positive for *E. coli* correctly predicted, and specificity is the fraction of cases with *E. coli* not detected correctly predicted. For this comparison, the models all include the same explanatory variables (i.e., no reduced form of the model) and the same number of observations (i.e., sample data used to construct the modes, ‘n’).

Model	**Sensitivity**	**Specificity**	**% Correctly Predicted Overall**	**True Positives**	**True Negatives**
OLS (n=1080)	100%	0%	72%	774	306
BLR (n=1080)	98%	7%	73%	774	306
CT (n=1080)	94%	16%	72%	774	306

Table S6 Comparison of full (i.e., all variables and no missing observations) 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	**B**[d]	**SE**	**B**	**SE**	**B**	**SE**	**Prune Level**
Constant	1.9	0.20	0.25	0.61	1.20**	0.57	-
Respondent works outside the home[e]	-0.20***	0.20	-0.41**	0.19	-0.46**	0.19	4
Regular weekly expenditure *per capita*	1.1x10^6	0.00	0.00	0.00	0.00	0.00	-
House has dirt floor[e]	0.10	0.07	0.23	0.19	0.22	0.19	2
House located within town[e]	-0.10	0.07	-0.18	0.19	-0.31	0.19	-
Infant present in household[e]	0.08	0.06	0.13	0.18	0.35*	0.18	-
Household has private latrine[e]	0.07	0.06	0.37**	0.17	0.20	0.17	3
Feces visible around household[e]	0.161	0.12	0.29	0.37	0.57	0.36	-
Latrine has a cement floor[e]	-0.07	0.08	0.08	0.20	-0.14	0.21	3
Children open defecate[e]	0.02	0.06	-0.01	0.17	0.13	0.17	-
Water source is improved[e]	-0.54***	0.08	-1.00***	0.28	-1.57***	0.27	5
Water was actively treated[e]	0.07	0.08	-0.15	0.23	0.24	0.22	-
Water extracted in risky manner[e]	-0.03	0.09	0.43*	0.26	0.04	0.23	-
Hand contacted water when extracting[e]	0.13	0.08	0.21	0.23	0.49**	0.23	-
Water stored for less than 24h[e]	-0.14**	0.07	-0.43**	0.19	-0.24	0.19	-
Log EC CFU/100mL on hands of caregiver	0.07**	0.03	0.06	0.08	0.17**	0.08	2
Someone in household has GI illness[e]	0.05	0.10	-0.33	0.31	0.27	0.27	-
Latrine has a roof[e]	-0.05	0.07	0.06	0.19	-0.14	0.20	-
Latrine has a septic tank[e]	-0.06	0.12	-0.22	0.31	-0.43	0.34	-
Latrine has a pit cover[e]	-0.06	0.069	-0.15	0.19	-0.15	0.19	-
Flies present in latrine[e]	-0.05	0.07	0.01	0.19	-0.106	0.191	-
Water storage container covered[e]	-0.07	0.14	0.34	0.41	-0.10	0.37	-
Water source on-plot[e]	0.01	0.09	0.12	0.23	-0.04	0.25	-

[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] Unstandardized Beta coefficient. [e] Binary variable (0 or 1)
***p<0.01 **0.01≥p<0.05 *0.05≥p<0.10

Table S7 Comparison of full (i.e., all variables and no missing observations) 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 [d]**	**SE**	**B**	**SE**	**Prune Level**
Constant	2.36***	0.14	0.71**	.33	-
Respondent works outside the home [e]	0.15	0.08	0.44**	.18	-
Regular weekly expenditure *per capita*[g]	-0.01*	0.01	0.00	0.017	-
House located in town [e]	0.40***	0.07	0.54***	0.16	1
Infant present in household [e]	0.18**	0.07	0.12	0.16	-
Household has private latrine [e]	-0.21***	0.07	-0.39**	0.16	-
Feces visible around household [e]	0.29**	0.13	0.89**	0.36	-
Latrine has a cement floor [e]	-0.02	0.08	0.03	0.19	1
Latrine has a septic tank [e]	-0.20	0.13	0.04	0.29	-
Flies present in latrine [e]	0.04	0.07	-0.29*	0.17	1
Children open defecate [e]	0.06	0.07	0.24	0.15	-
Time since last hand washing 1h or less [e]	-0.08	0.07	-0.12	0.16	-
Prior activity involved washing [e,f]	0.21**	0.09	0.49**	0.22	-
Prior activity food handling [e,f]	0.12	0.08	0.40**	0.19	-
Prior activity 'other' [e,f]	0.11	0.13	0.39	0.32	-
Prior activity (for classification tree ONLY) [g]	-	-	-	-	-
Someone in household has GI illness [e]	0.04	0.11	0.35	0.28	-
House has a dirt floor [e]	-0.-3	0.07	-0.08	0.17	-
Latrine has a roof [e]	-0.04	0.08	-0.11	0.17	-
Latrine has a pit cover [e]	0.10	0.08	0.23	0.17	-
Respondent has primary education [e]	-0.08	0.07	-0.26	0.17	-
Hand washing station with soap present [e]	-0.07	0.08	-0.07	0.17	-
Hands dried with fabric after hand washing [e]	-0.05	0.07	0.01	0.15	-
Hand wetted for hand washing by pouring water [e]	-0.06	0.09	-0.07	0.20	-

[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] Unstandardized Beta coefficient [e] Binary variable (0 or 1) [g] In (1000 Tsh) [h] Dummy variables with the reference activity of 'sitting' [g] Categorical variable of activity prior to hand rinse being sitting, washing, food handling, or other

***p<0.01 **0.01≥p<0.05 *0.05≥p<0.10