Supporting Information

Improved decision making for water lead testing in U.S. child care facilities using machine-learned Bayesian networks

Riley E. Mulhern,[1*] AJ Kondash,[1] Ed Norman,[2] Joseph Johnson,[1] Keith Levine,[1] Andrea McWilliams,[1] Melanie Napier,[2] Frank Weber,[1] Laurie Stella,[1] Erica Wood,[1] Crystal Lee Pow Jackson,[1] Sarah Colley,[1] Jamie Cajka,[1] Jackie MacDonald Gibson,[3] Jennifer Hoponick Redmon[1*]

1 – RTI International, Research Triangle Park, NC, 27709, USA
2 – Environmental Health Section, Division of Public Health, North Carolina Department of Health and Human Services, Raleigh, NC, 27609, USA
3 – North Carolina State University, Department of Civil, Construction, and Environmental Engineering, Raleigh, NC, 27695, USA

*Corresponding authors: Riley E. Mulhern (rmulhern@rti.org) and Jennifer Hoponick Redmon (jredmon@rti.org)

CONTENTS

Additional Methods Information 6

Supporting Figures

Figure S1. *Flow chart of model development and performance evaluation steps.* 4

Figure S2. *Histograms of building-wide maximum and 90th percentile lead concentrations across all centers.*...........5

Figure S3. *Comparison of model performance after 10-fold cross validation using various variable selection criteria, including AIC, BIC, Chi-squared tests, and bootstrapping. The various criteria yielded similarly performing models except for BIC. For this comparison, missing values in the training set were imputed using structural expectation maximization. Chi-squared tests were ultimately selected as the variable selection criteria for their ability to handle missing values.*...7

Figure S4. *Strength of association between the target and all possible predictor nodes for model Max>=1 determined by Chi-Squared tests. Predictors not directly associated with the target node (dashed lines) were removed from the model.* 7

Figure S5. *Missingness pattern of the complete data set. Overall, the data set contained only 6.2% missing values, but over half of the facilities were missing data for at least one field.*..8

Figure S6. *Comparison of model performance after 10-fold cross validation using various approaches to handle missing values in the data set, including using complete observations from each nodes local network, multiple imputation by chained equations (MICE) and structural expectation maximization (SEM). Overall, the results were not significantly affected by the imputation procedure.* 10

Figure S7. *Structures of the eight final models. Interactive versions can also be seen at:* *www.cleanwaterforcarolinakids.org/publications/bn_models*..11

Figure S8. *Receiver operating characteristic (ROC) curves for each model. Models predicting the maximum lead level (max >1, >5, >10, and >15) are shown on the left. Models predicting the 90th percentile lead level (P90 >1, >5, >10, >15) are shown on the right.*..12

Figure S9. *Relationship between the prior probability of each model, indicating the level of class imbalance, and the AU-ROC (pink), AU-PR (grey), and F2-score (gold) performance metrics (F1-scores are not shown for visual clarity, but follow a similar trend as F2-scores). AU-ROC values are insensitive to large class imbalance, while AU-PR and F2-scores reveal decreasing predictive performance for more rare outcomes.*..13

Figure S10. *Precision-recall (PR) curves for each model. Models predicting the maximum lead level (max >1, >5, >10, and >15) are shown on the left. Models predicting the 90th percentile lead level (P90 >1, >5, >10, >15) are shown on the right.*...14

Figure S11. *Posterior probabilities of the four most frequently selected variables across all eight models compared to each model's prior. The width of the bars shows the range of impact of each variable on building-wide lead risk. Grey diamonds indicate the prior probability of each model target considering all states. Models are ordered on the y-axis according to the size of the variable's impact on the target, with models where the variable was more impactful at top.* 15

Figure S12. *Posterior probabilities for the variable "Past Faucet Fixture Change" for the seven models where it was included. If the center did not know whether any past faucets had been changed significantly increased the building-wide water lead risk.* 16

Figure S13. *Posterior probabilities for the variables "Home-based" (Panel A) and "School-based" (Panel B) highlighting decreases in risk for small home-based centers, and increases in risk for larger school-based centers for certain model targets.* 16

Figure S14. *Posterior probabilities for the variables "Phosphate addition" (Panel A) and "pH adjustment" (Panel B) highlighting increases in water lead risk in child care centers served by water systems (public or private) that do not use phosphate-based corrosion inhibitors or pH adjustment.* 17

Figure S15. *Posterior probabilities for the variable "Past LCR exceedance" for the five models where it was included. If the center relied on a private well it was coded as "LCR NA," indicating that Lead and Copper Rule monitoring is not applicable. Private well water significantly increased the building-wide water lead risk, whereas a past action level (AL) exceedance by the water utility decreased the risk.* 17

Figure S16. *Tornado charts showing the minimum and maximum posterior probabilities and node states for all variables included in each model. The dashed vertical line in each chart shows the prior probability of the target for each model. Posterior probabilities to the right indicate increased risk; posterior probabilities to the left indicate decreased risk. The width of the line indicates the magnitude of the effect of the variable on the target.* 18

Figure S17. *Plots comparing the sensitivity improvement and sampling reduction achieved by each model compared to the alternative heuristics. Models predicting the maximum lead level (max >1, >5, >10, and >15) are shown on the left. Models predicting the 90th percentile lead level (P90 >1, >5, >10, >15) are shown on the right.* ...19

Supporting Tables

Table S1. *Variables included in the Clean Water for Carolina Kids data set[16] used as predictors in the BN models to predict building-wide lead risk.* 20

Table S2. *Additional variables included in the data set compiled from publicly available data sources.* 22

Table S3. *Summary of all performance metrics for each model.* 24

Table S4. *Summary of mean F1 and F2 scores for each heuristic from 10-fold cross validation.* 24

Table S5. *Summary table of all significant variables included in each model. Interactive model structures can also be seen at www.cleanwaterforcarolinakids.org/publications/bn_models.* 25

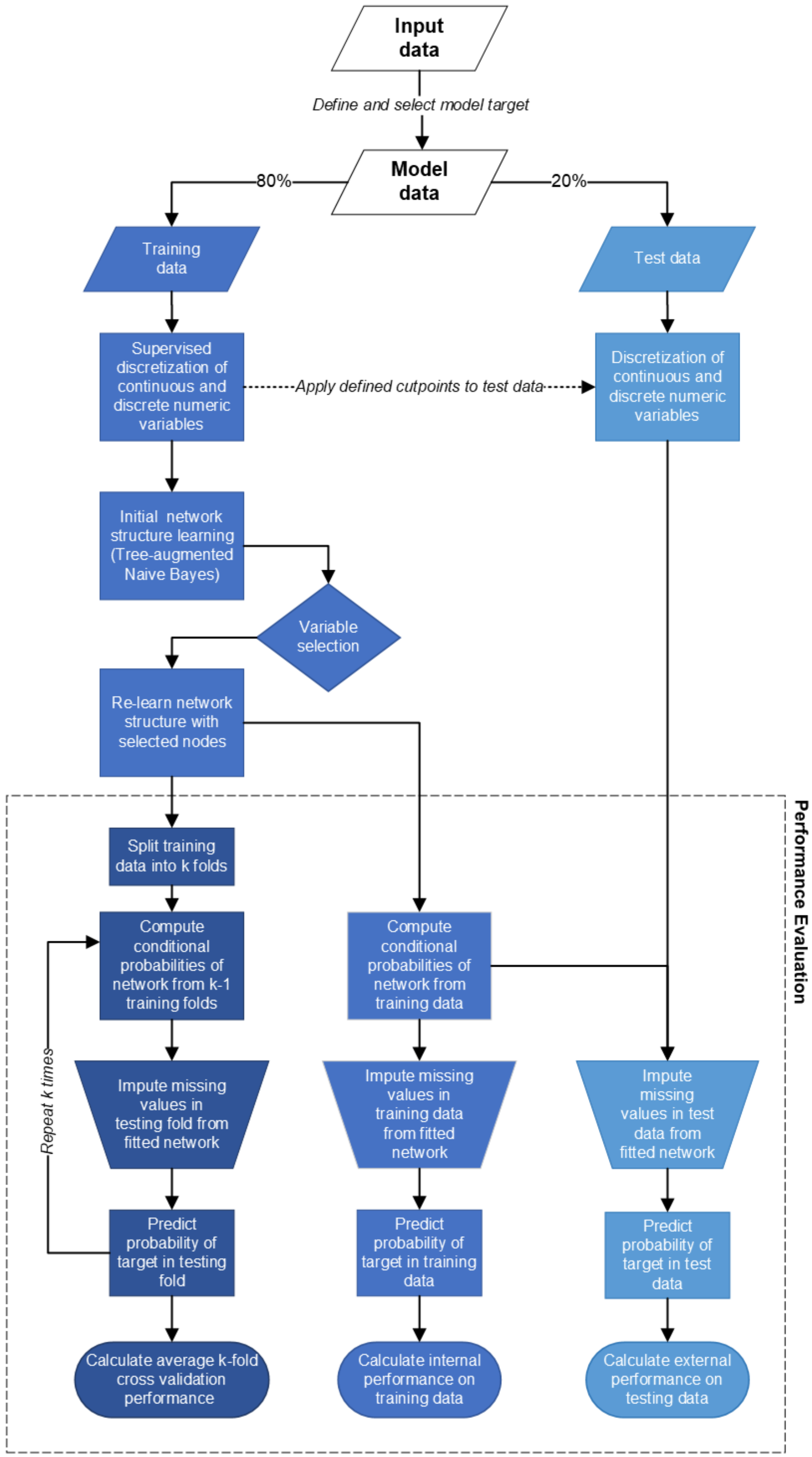

Figure S1. Flow chart of model development and performance evaluation steps.

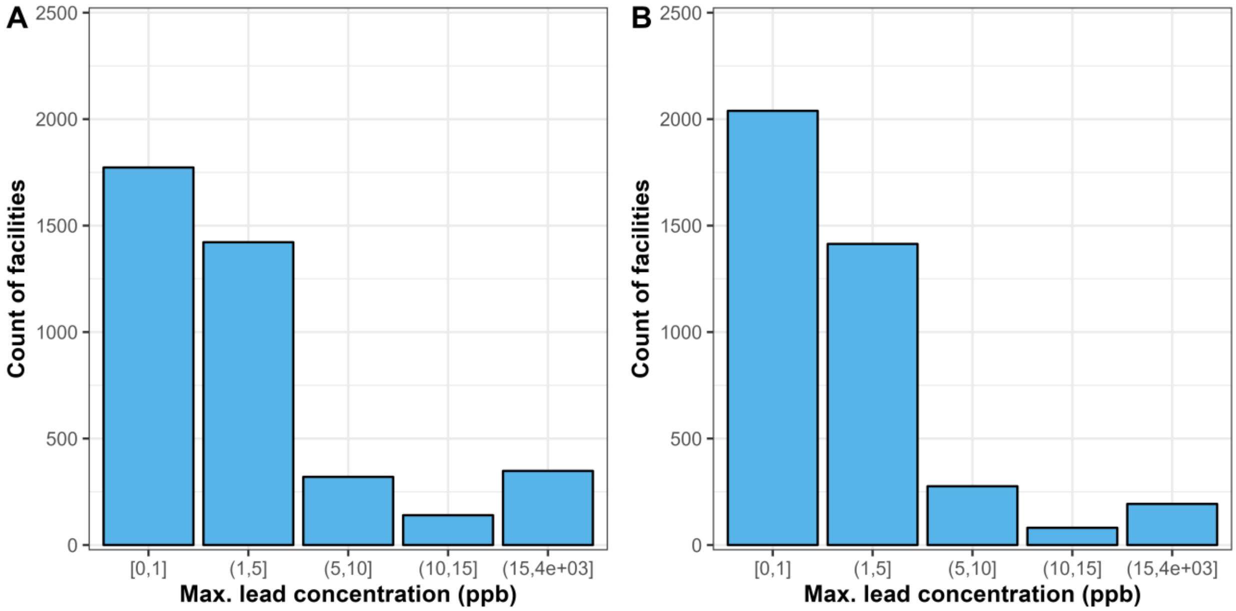

Figure S2. Histograms of building-wide maximum and 90th percentile lead concentrations across all centers.

Additional Methods Information

Discretization

Several other supervised discretization algorithms were also tested, such as Minimum Description Length Principle (MDLP), Class-Attribute Interdependence Maximization (CAIM), and Ameva algorithms (each of which has been shown to adequately represent the true data distribution with a minimal number of discrete intervals[1–3]), but these were found to inadvertently remove certain variables before the machine learning step by failing to select any cut points, thus rendering them unusable in a BN model. Since variables in both the training and test set must have the same intervals for BN models, the cut points selected by the discretization algorithm for the training set were then manually applied to the test set. Although discretization of the complete data set prior to machine learning (rather than performing separate discretization steps on the training and test sets) is common in BN modelling,[4–6] the manual application of the training set cut points to the test set minimizes potential data leakage concerns.

Structure learning and feature selection

Additional measures of arc strength were assessed alongside Chi-squared tests, including whether removing the node from the network would result in a reduction of the model's Akaike Information Criterion or Bayesian Information Criterion, and whether the node was connected to the target at least 10% of the time during 100 bootstrapped unsupervised structure learning iterations. The various variable selection approaches performed comparably with the exception of the Bayesian Information Criterion (**Figure S3**). Chi-squared tests were chosen because they allowed missing values in the training set to be handled transparently, whereas the other approaches required imputation as a pre-processing step (see "Missing values processing" below). An example of the arc strengths calculated by the Chi-squared tests can be seen for the Max≥1 model in **Figure S4**. The bootstrapping procedure yielded significantly fewer variables and could be used to simplify the models further if many of the variables selected by Chi-squared tests were unavailable in certain areas.

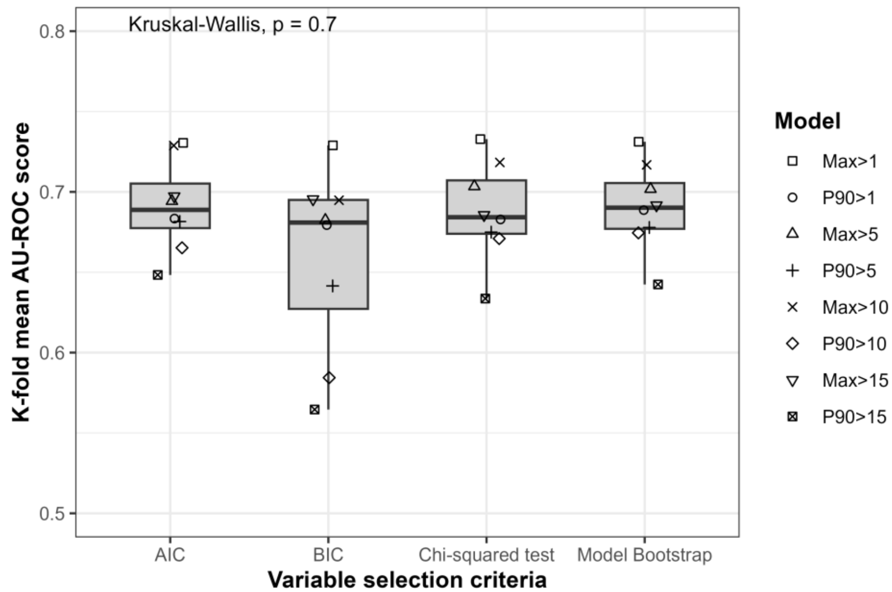

Figure S3. Comparison of model performance after 10-fold cross validation using various variable selection criteria, including AIC, BIC, Chi-squared tests, and bootstrapping. The various criteria yielded similarly performing models except for BIC. For this comparison, missing values in the training set were imputed using structural expectation maximization. Chi-squared tests were ultimately selected as the variable selection criteria for their ability to handle missing values.

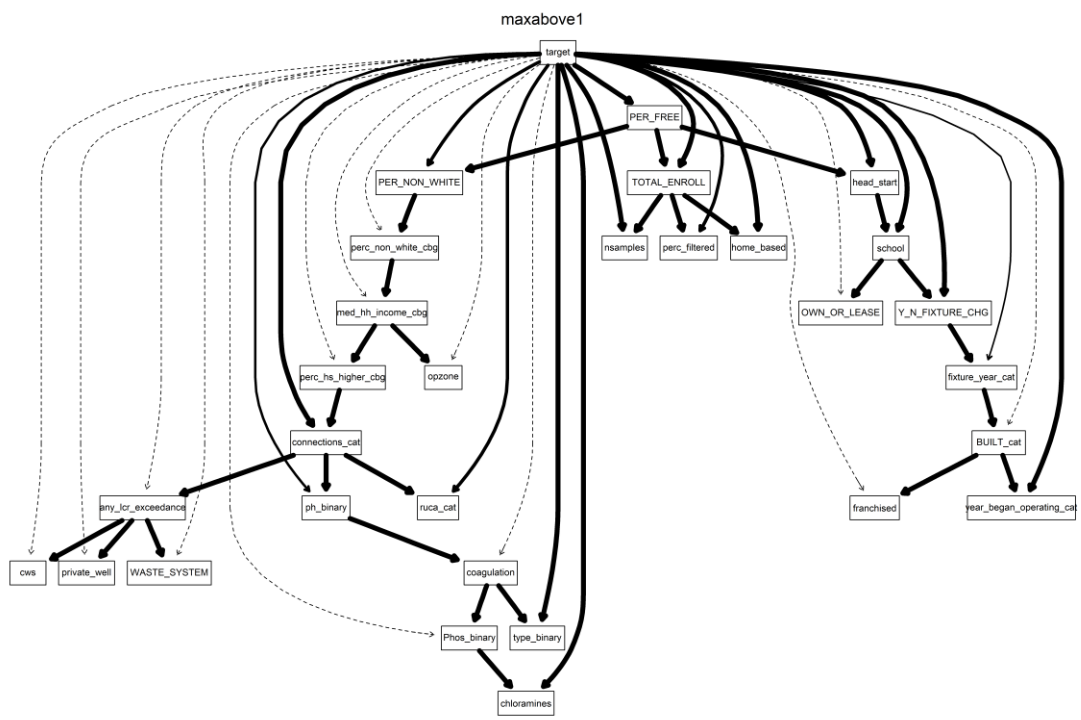

Figure S4. Strength of association between the target and all possible predictor nodes for model Max>=1 determined by Chi-Squared tests. Predictors not directly associated with the target node (dashed lines) were removed from the model.

Missing values processing

The final data set contained only 6.2% missing values. While there is no recognized cutoff for an acceptable level of missingness for statistical analysis, <10% missing data is generally considered low, with little bias expected on the result.[7] However, a little over half of the facilities were missing data for at least one field. The missingness pattern can be seen in **Figure S5**.

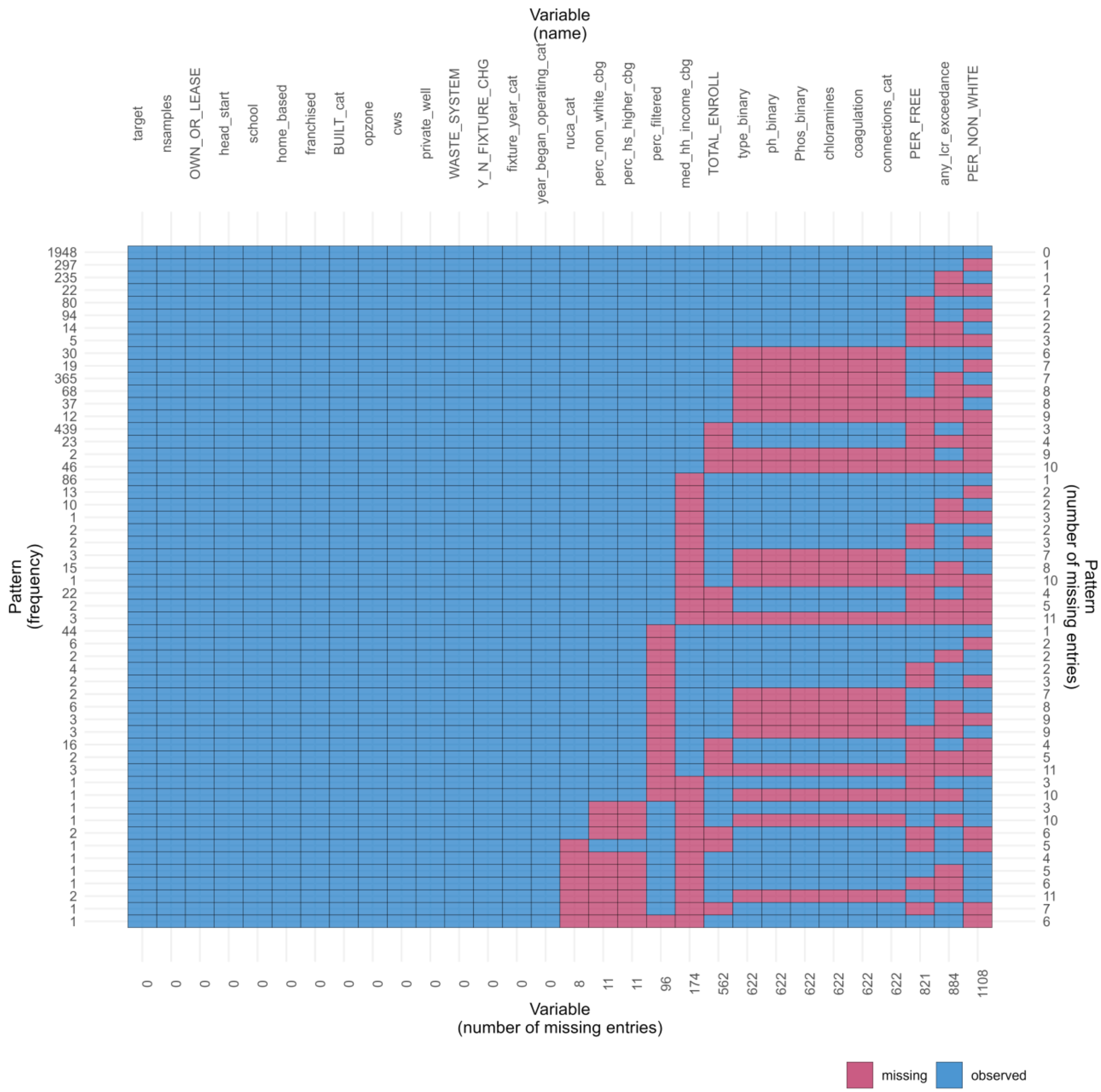

Figure S5. Missingness pattern of the complete data set. Overall, the data set contained only 6.2% missing values, but over half of the facilities were missing data for at least one field.

Most of the missing data were related to enrollment and demographic information (e.g., percent non-white students, percent free/reduced lunch students, total enrollment) that was not entered or was entered incorrectly by the facility during their registration with the Clean Water for Carolina Kids program (for example, erroneous enrollment numbers resulting in >100% non-White; in these cases, the values were treated as missing). These data are considered missing not at random (MNAR) since the missingness mechanism is likely a function of the variable itself (e.g., centers with a higher proportion of children of color may be more likely to accidently misreport their student demographic information compared to centers with only White students, or centers with a higher proportion of free/reduced lunch children may not want this information to be known due to stigma around receiving free services[8]). Additional missing data were related to Lead and Copper Rule monitoring and water system information in cases where facility staff either did not know or did not report which community water system they were served by, were served by a private well, or operated as a NTNC water system. We consider these data to be missing at random (MAR) because their missingness depends on other factors in the data set, such as water system type.

To assess the effect of missingness in the training data set on model development, we conducted a sensitivity analysis to evaluate several different approaches to handling missing values,[9] including multiple imputation by chained equations (MICE),[10,11] structural expectation maximization (SEM) from a separate unsupervised BN,[11,12] estimating conditional probability distributions for each network node using complete observations for that node's local network.[13] The results of this sensitivity analysis can be seen in **Figure S6**. Overall, the MI and SEM imputation approaches provided negligible benefit in performance while significantly increasing computational costs compared to handling missing values transparently.

Meanwhile, missing values could not be ignored in the test data set since the "predict" function in *bnlearn* requires complete evidence to predict the probability of the target node. One advantage of BNs is that, once trained, the network structure itself can be used to estimate missing values in a test set using Bayesian likelihood weighting.[14] Given the potential difficulty of collecting building-specific data in some cases, missing values are highly likely in practice. Thus, practitioners could feasibly apply our models to new data without having to develop a separate imputation procedure to handle missing values before prediction. To eliminate data leakage concerns and to best simulate the model's practical performance where no prior knowledge of the outcome is available, our model strips the test data (whether the external test set, the test fold during k-fold cross validation, or the training data itself when assessing the internal model performance) of the real result of the target outcome and generates a random guess in its place. A random guess is necessary because the “impute” function in *bnlearn* needs a data set that matches the structure of the network, so the target variable cannot be blank. Importantly, this simulates how practitioners would approach the model with missing data and does not introduce any “unfair” information to the model during prediction. The random guess is then removed after the imputation step and the predicted probability of the target node is calculated using the imputed testing data as evidence.

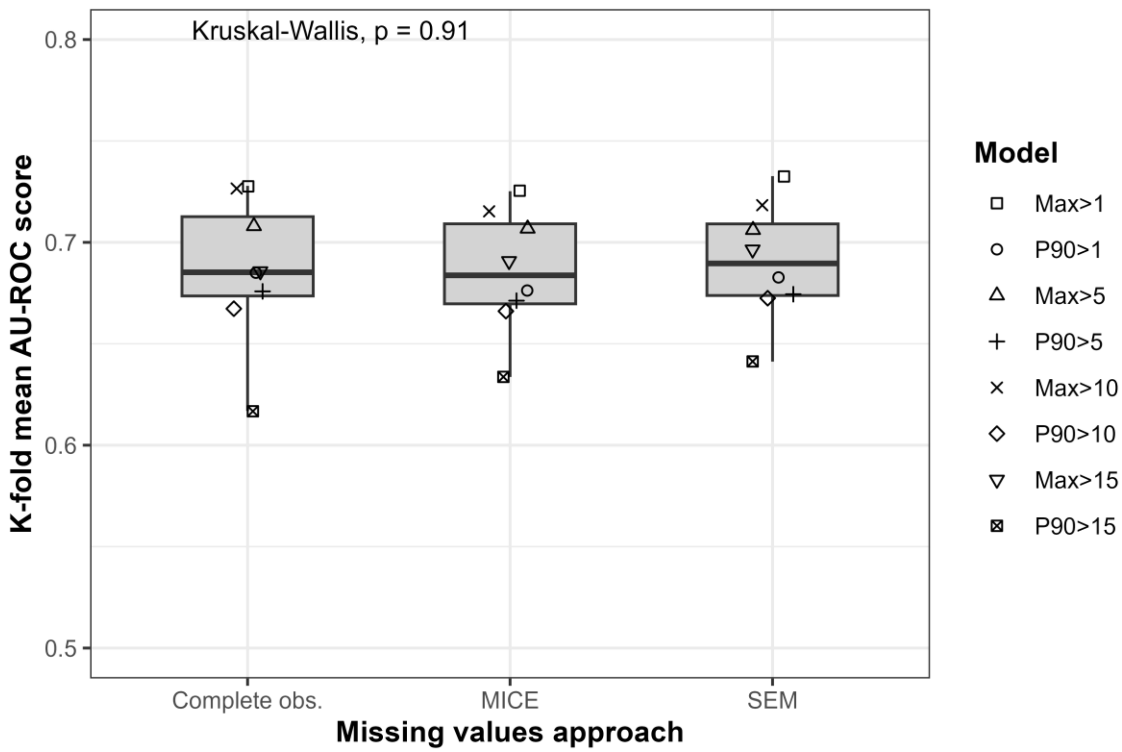

Figure S6. Comparison of model performance after 10-fold cross validation using various approaches to handle missing values in the data set, including using complete observations from each nodes local network, multiple imputation by chained equations (MICE) and structural expectation maximization (SEM). Overall, the results were not significantly affected by the imputation procedure.

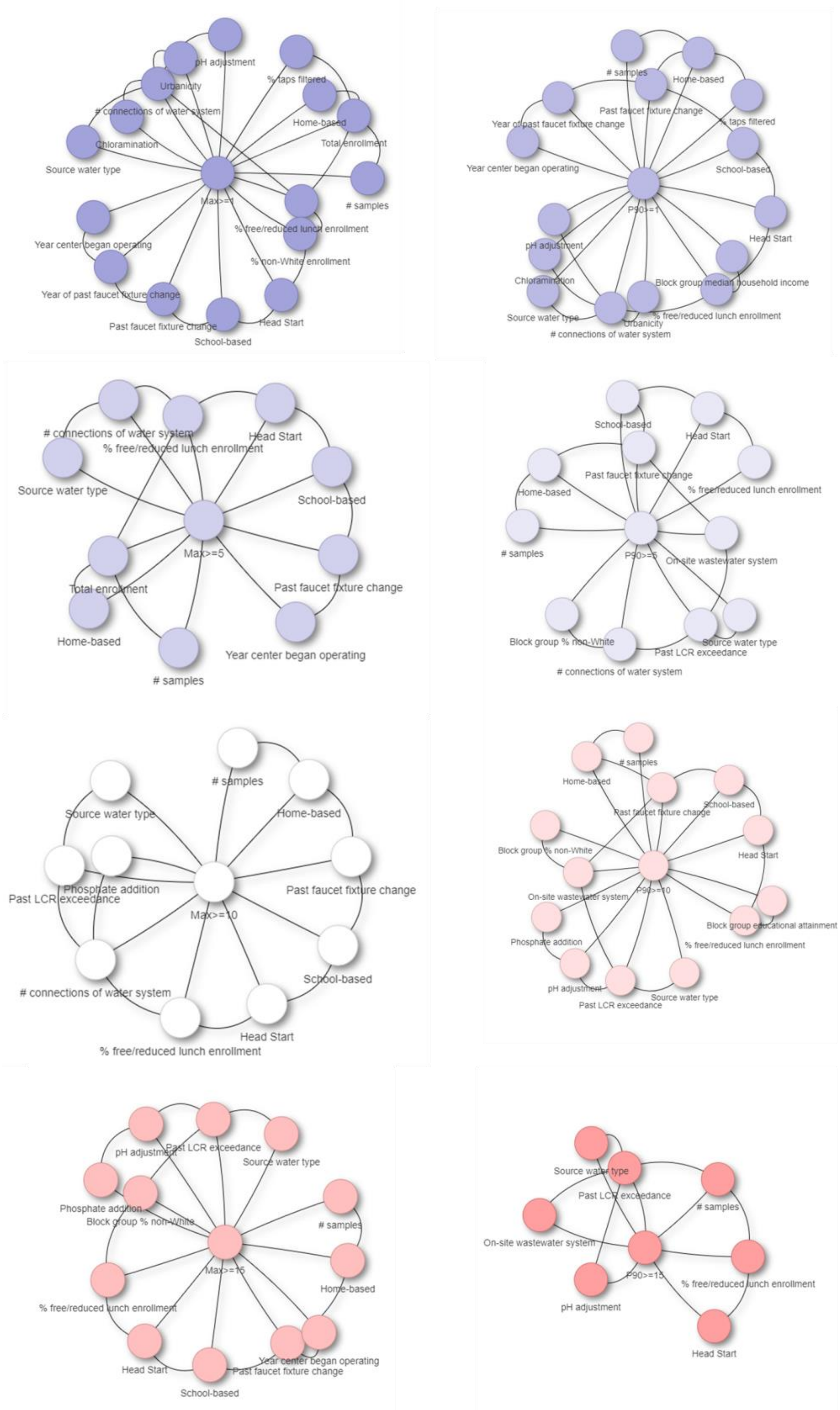

Figure S7. Structures of the eight final models. Interactive versions can also be seen at: www.cleanwaterforcarolinakids.org/publications/bn_models

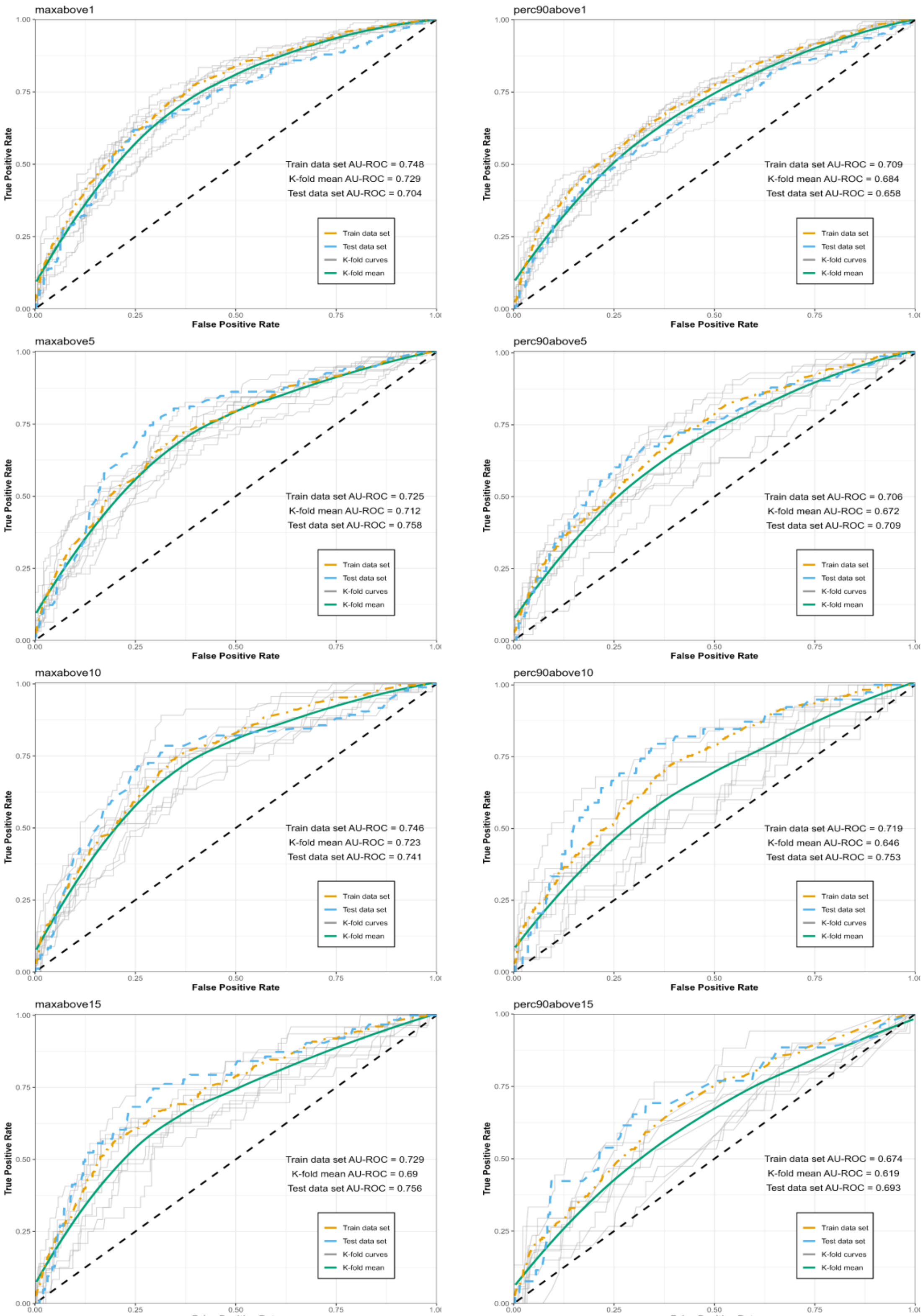

Figure S8. Receiver operating characteristic (ROC) curves for each model. Models predicting the maximum lead level (max >1, >5, >10, and >15) are shown on the left. Models predicting the 90th percentile lead level (P90 >1, >5, >10, >15) are shown on the right.

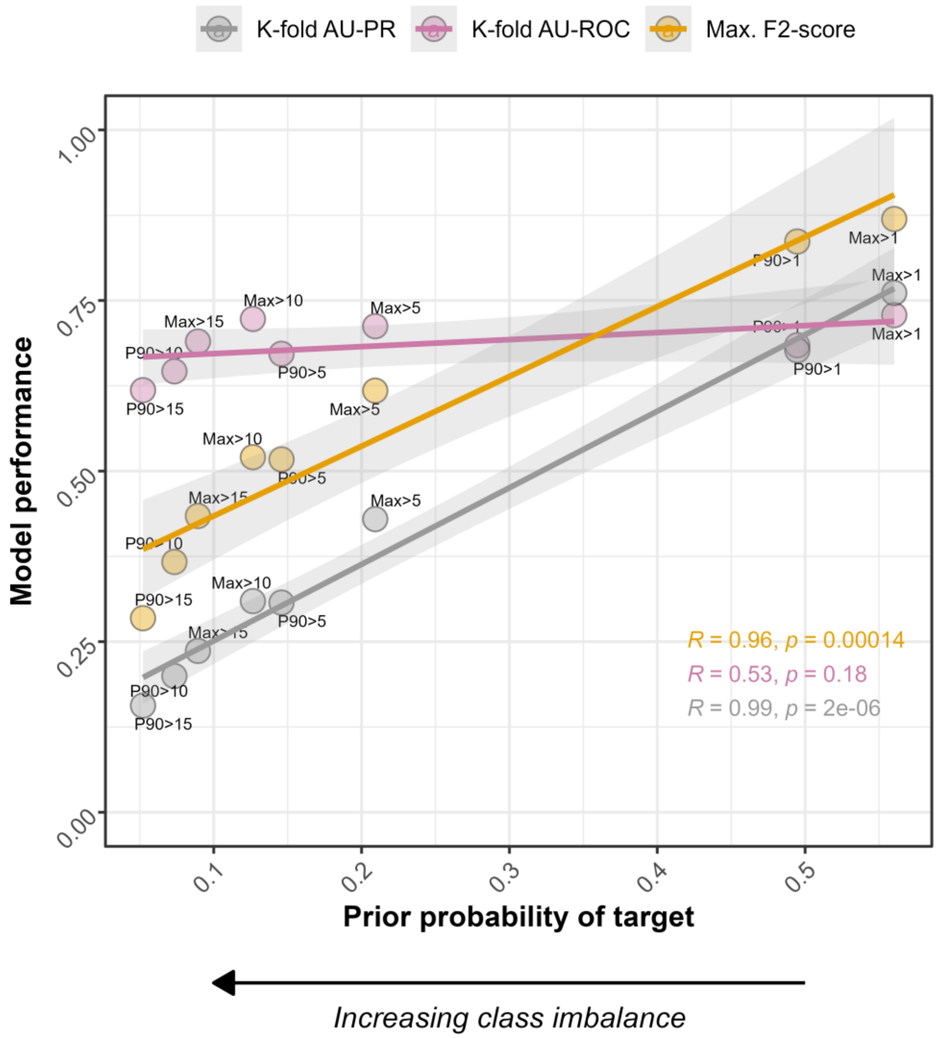

Figure S9. Relationship between the prior probability of each model, indicating the level of class imbalance, and the AU-ROC (pink), AU-PR (grey), and F2-score (gold) performance metrics (F1-scores are not shown for visual clarity, but follow a similar trend as F2-scores). AU-ROC values are insensitive to large class imbalance, while AU-PR and F2-scores reveal decreasing predictive performance for more rare outcomes.

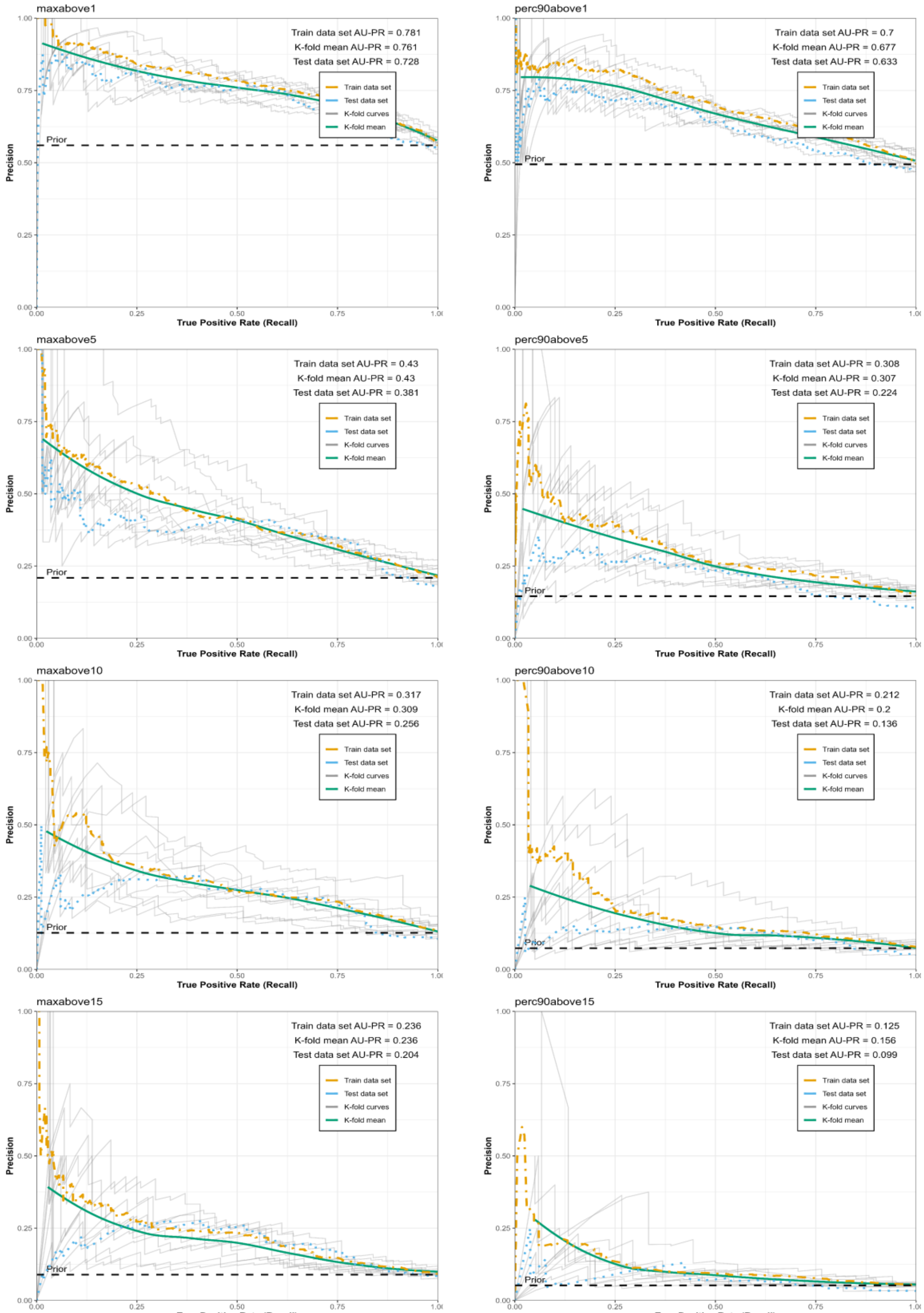

Figure S10. Precision-recall (PR) curves for each model. Models predicting the maximum lead level (max >1, >5, >10, and >15) are shown on the left. Models predicting the 90th percentile lead level (P90 >1, >5, >10, >15) are shown on the right.

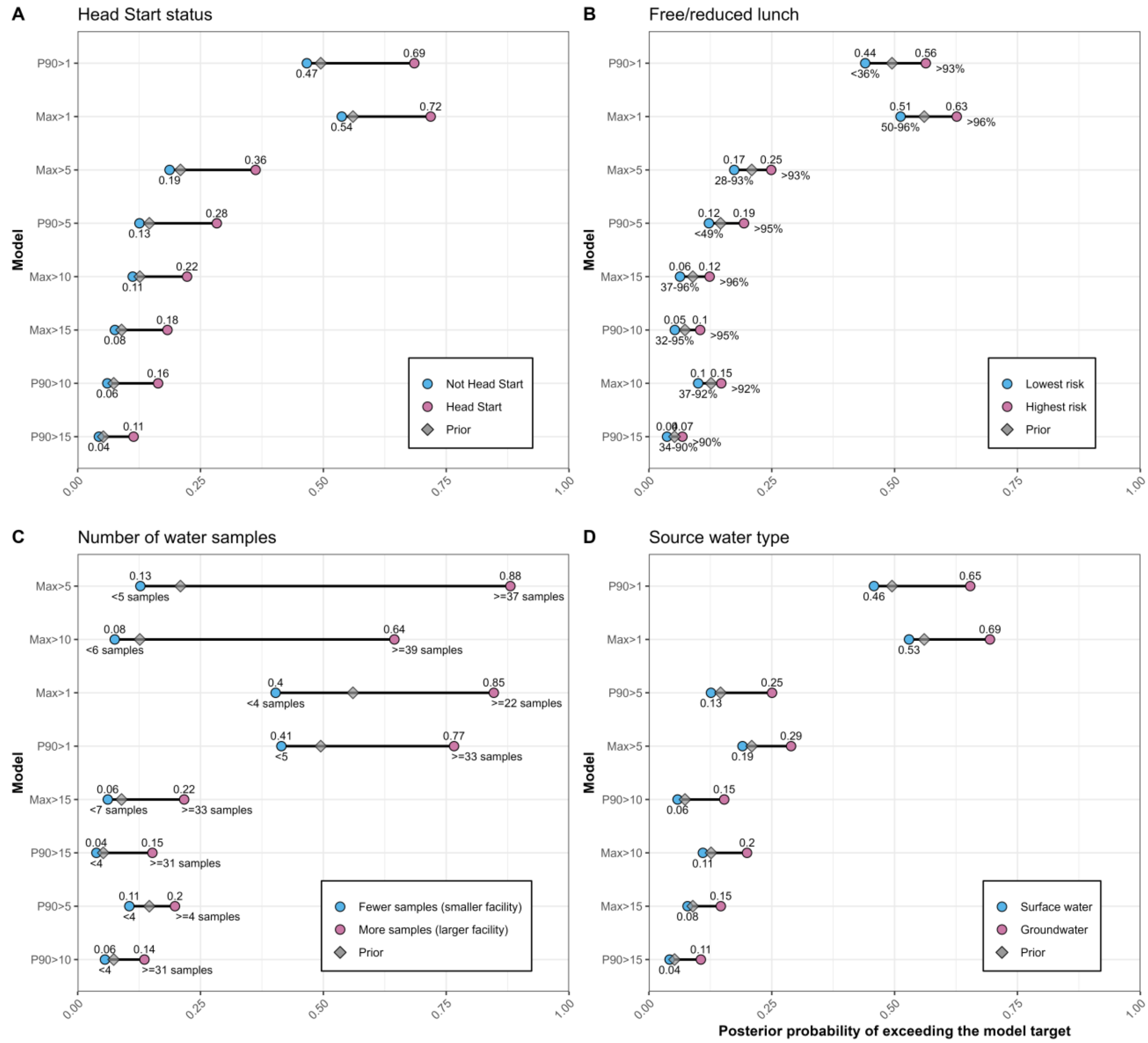

Figure S11. Posterior probabilities of the four most frequently selected variables across all eight models compared to each model's prior. The width of the bars shows the range of impact of each variable on building-wide lead risk. Grey diamonds indicate the prior probability of each model target considering all states. Models are ordered on the y-axis according to the size of the variable's impact on the target, with models where the variable was more impactful at top.

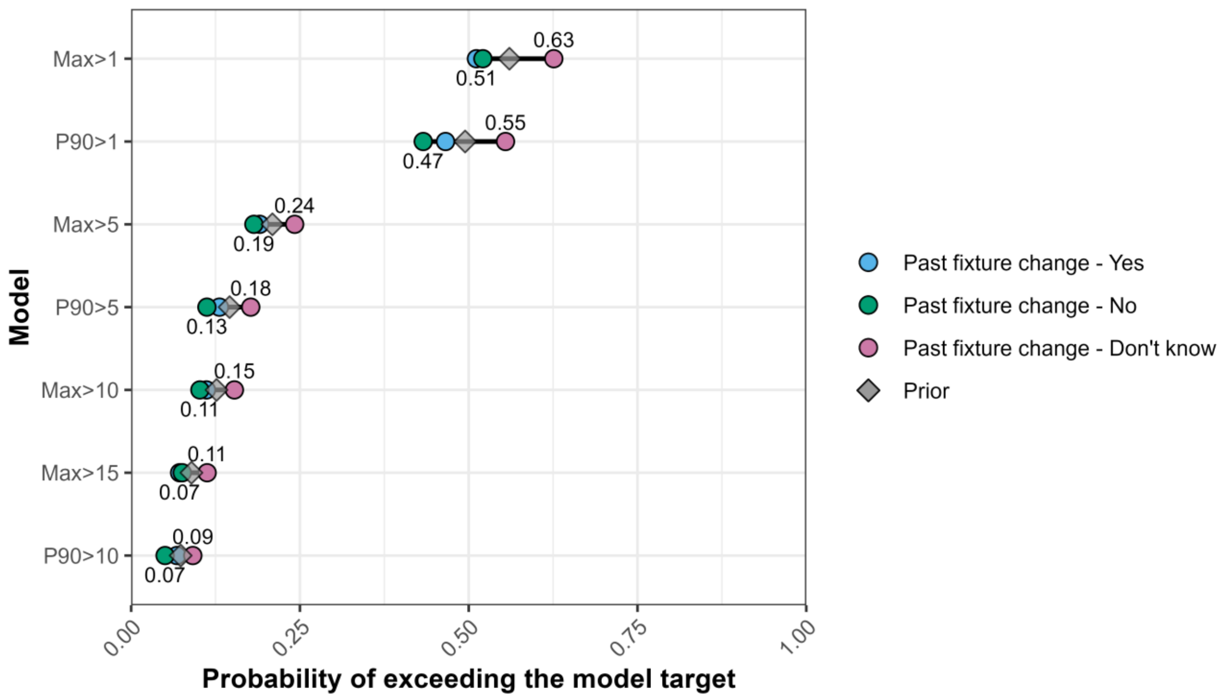

Figure S12. Posterior probabilities for the variable "Past Faucet Fixture Change" for the seven models where it was included. If the center did not know whether any past faucets had been changed significantly increased the building-wide water lead risk.

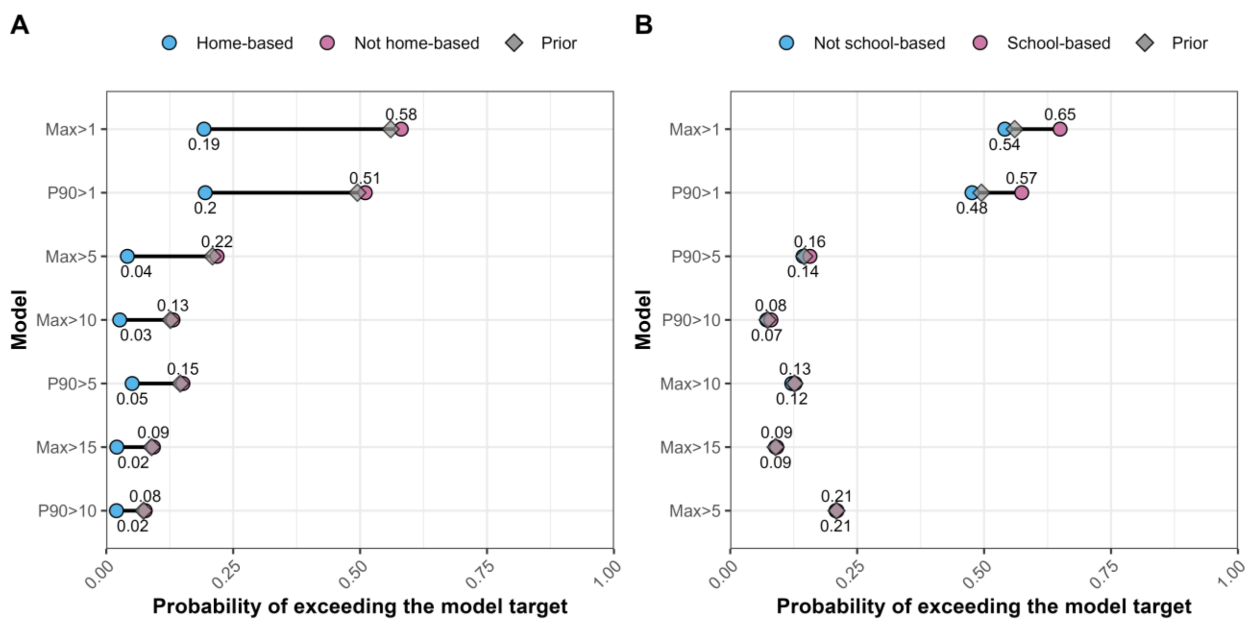

Figure S13. Posterior probabilities for the variables "Home-based" (Panel A) and "School-based" (Panel B) highlighting decreases in risk for small home-based centers, and increases in risk for larger school-based centers for certain model targets.

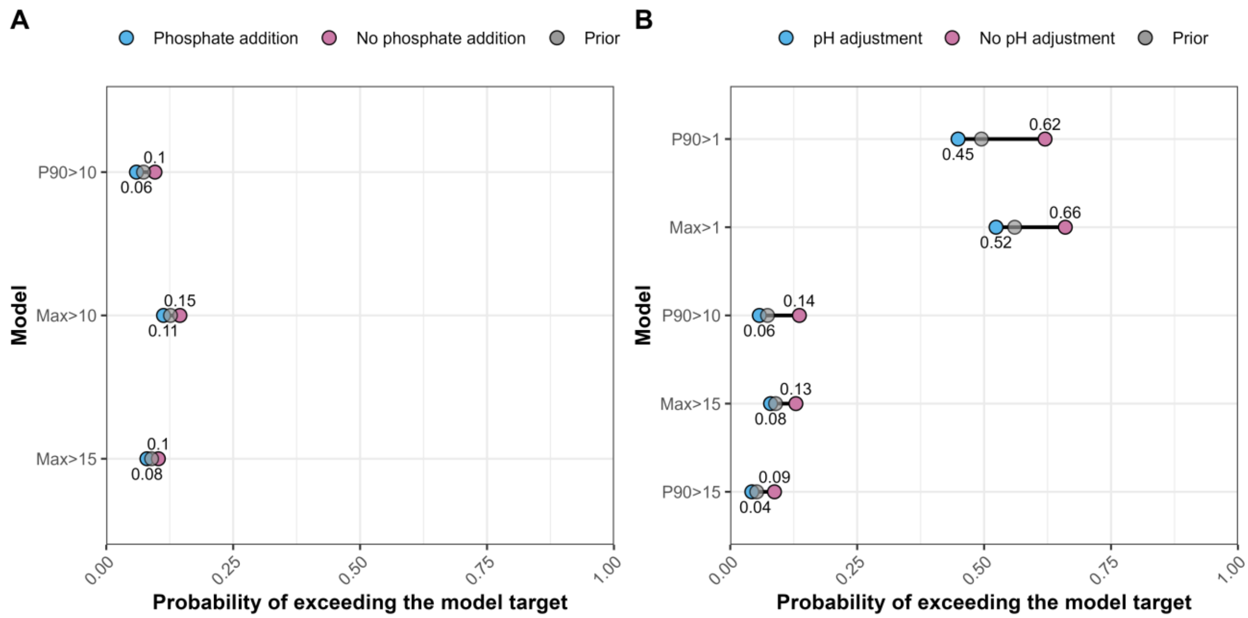

Figure S14. Posterior probabilities for the variables "Phosphate addition" (Panel A) and "pH adjustment" (Panel B) highlighting increases in water lead risk in child care centers served by water systems (public or private) that do not use phosphate-based corrosion inhibitors or pH adjustment.

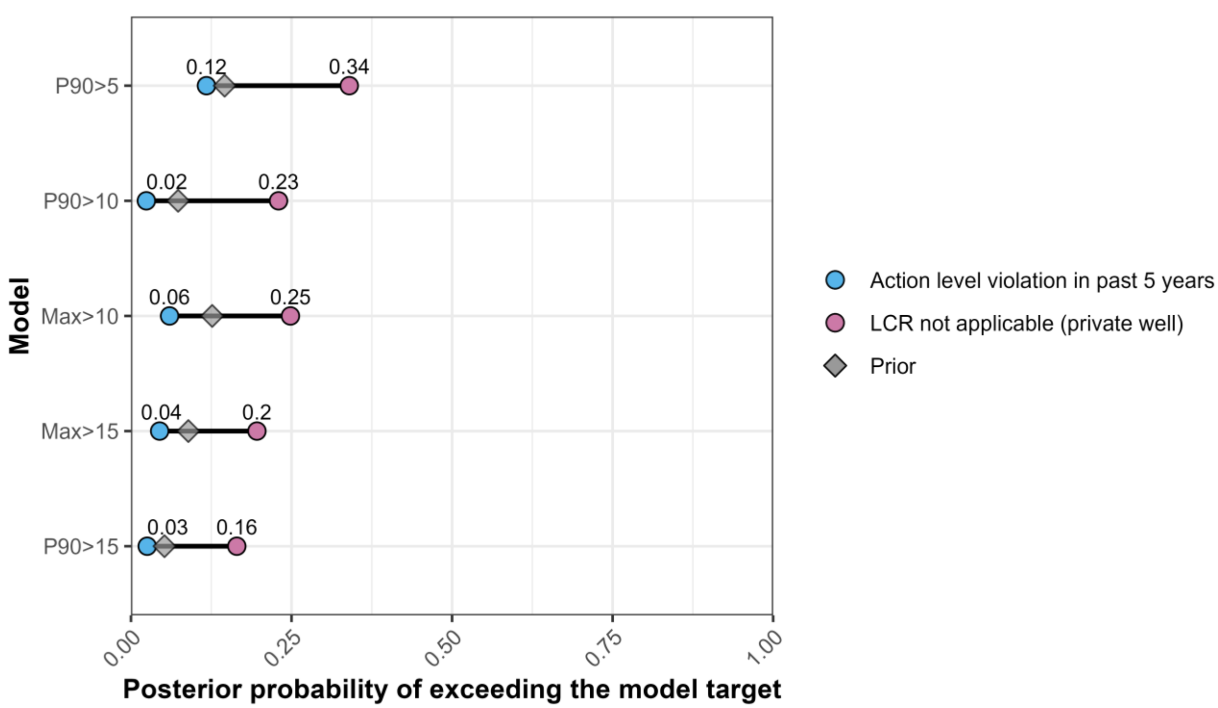

Figure S15. Posterior probabilities for the variable "Past LCR exceedance" for the five models where it was included. If the center relied on a private well it was coded as "LCR NA," indicating that Lead and Copper Rule monitoring is not applicable. Private well water significantly increased the building-wide water lead risk, whereas a past action level (AL) exceedance by the water utility decreased the risk.

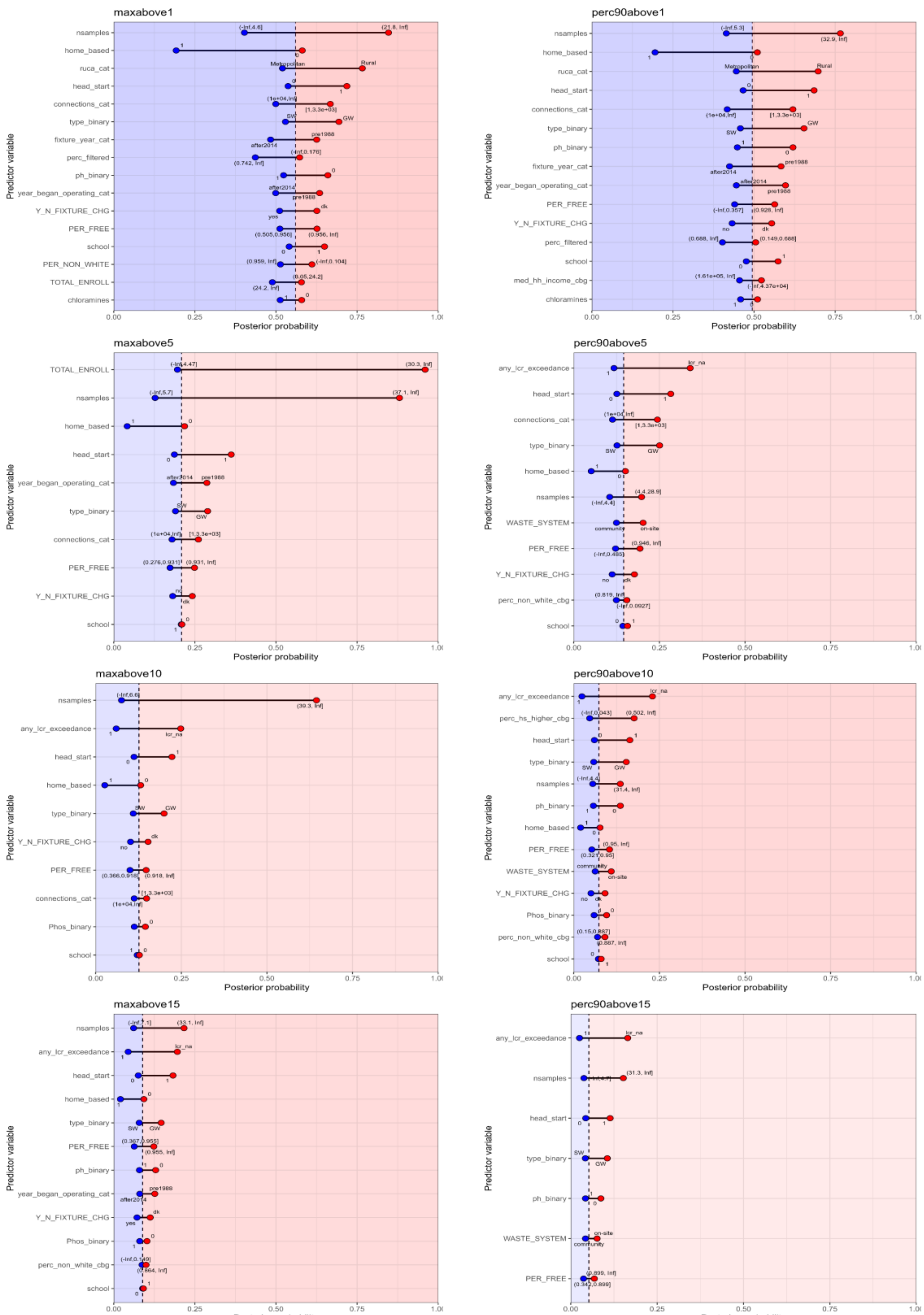

Figure S16. Tornado charts showing the minimum and maximum posterior probabilities and node states for all variables included in each model. The dashed vertical line in each chart shows the prior probability of the target for each model. Posterior probabilities to the right indicate increased risk; posterior probabilities to the left indicate decreased risk. The width of the line indicates the magnitude of the effect of the variable on the target.

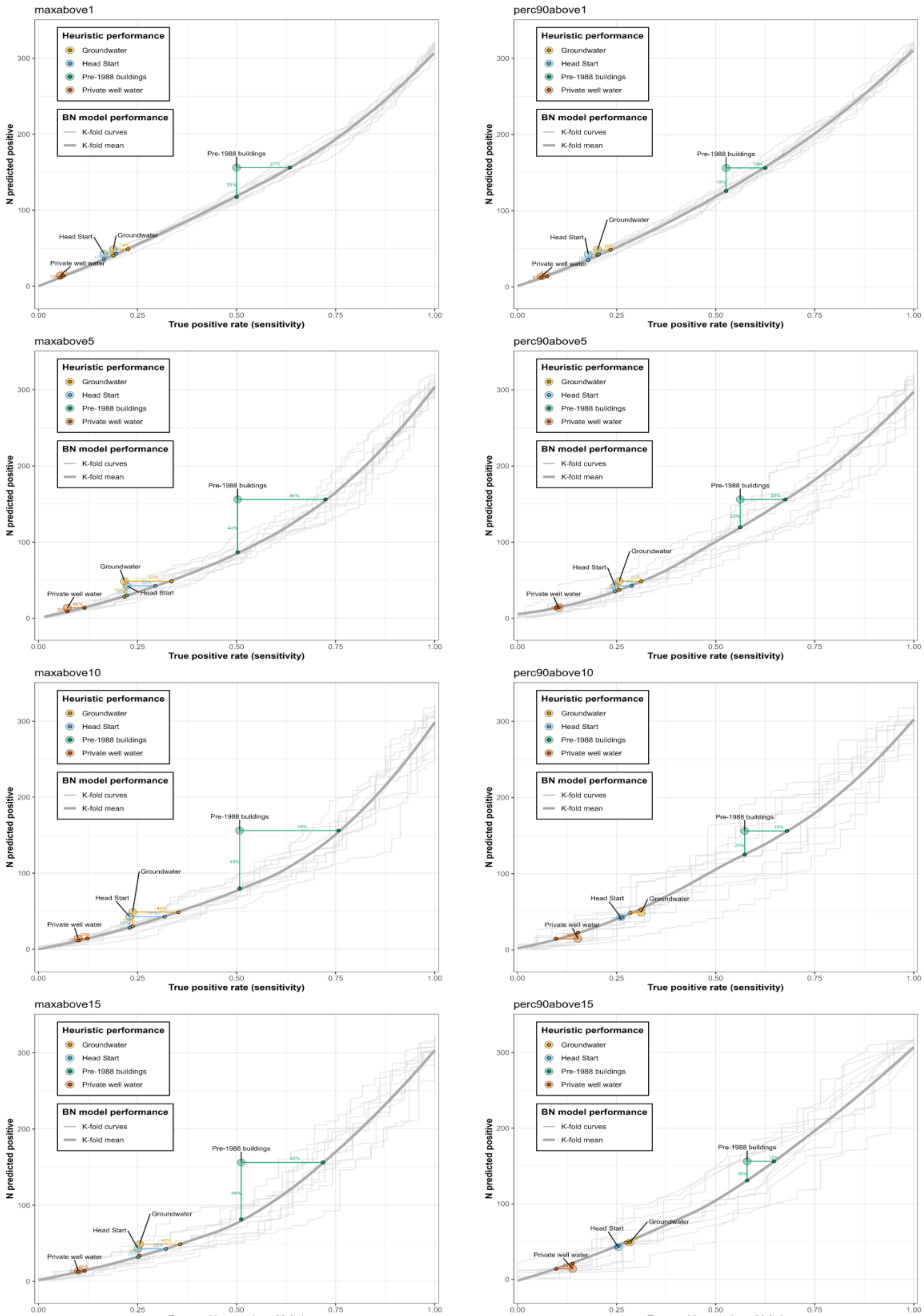

Figure S17. Plots comparing the sensitivity improvement and sampling reduction achieved by each model compared to the alternative heuristics. Models predicting the maximum lead level (max >1, >5, >10, and >15) are shown on the left. Models predicting the 90th percentile lead level (P90 >1, >5, >10, >15) are shown on the right.

Table S1. Variables included in the Clean Water for Carolina Kids data set[15] used as predictors in the BN models to predict building-wide lead risk.

Variable Name	Abbreviated name	Description	Variable type	Summary
Building information				
Ownership	OWN_OR_LEASE	Whether the center building was owned or leased.	Binary	Own: 2,774 Lease: 1,229
School-based	school	Whether the center was located in a school building	Binary	School-based: 742 Not school-based: 3,261
Home-based	home_based	Whether the center was located in a home	Binary	Home-based: 192 Not home-based: 3,811
Franchised	franchised	Whether the center was part of a franchise	Binary	Franchised: 205 Not franchised: 3,798
Year built	BUILT_cat	Year the building was constructed	Categorical (Pre-1988, 1988-2013, or Post 2014)	Pre-1988: 1,944 1988-2013: 1,875 After 2014: 184
Community water system	cws	Whether the center was connected to a public community water system (CWS)	Binary	Connected to CWS: 3,452 Not connected to CWS: 551
# samples	nsamples	The number of drinking and cooking water samples collected for analysis by the center (center staff were instructed to sample all drinking and cooking taps).	Discrete	Mean: 5.7 Median: 4 Max: 51 Missing: 0
Percent filtered	perc_filtered	The percentage of taps sampled by each center that had a point-of-use filter installed and was flagged as filtered by center staff.	Continuous	Mean: 12% Median: 0% Max: 100% Missing: 96 (2.4%)
Private well	private_well	Whether the center used a private well for its water supply. This was determined according to whether the center responded "No" to the question "Does your drinking water come from a public water treatment plant?" and "Yes" to the question "Does the center use well water?" This variable was not mutually exclusive with "Community water system"	Binary	Private well: 175 No private well: 3,828

		suggesting that a) some non-CWS systems receive water from a source other than a well, b) that center staff were not aware that they were served by a CWS, or c) the center staff were not aware that they received water from a private well.		
On-site wastewater system	WASTE_SYSTEM	Whether the center was served by an onsite wastewater system	Categorical (Community, On-site, Unknown)	Community: 2,580 On-site: 535 Unknown: 888
Past faucet fixture change	Y_N_FIXTURE_CHG	Whether the center had any known past faucet fixture changes	Categorical (Yes, No, Don't Know)	Yes: 1474 No: 876 Don't Know: 1,653
Year of past faucet fixture change	fixture_year_cat	Year of past faucet fixture change	Categorical (Pre-1988, 1988-2013, or Post 2014. If answer to "Fixture change" was "No," or "Don't Know" default to "Year built".)	Pre-1988: 1,013 1988-2013: 2,000 After 2014: 990
Year center began operating	year_began_operating_cat	Year the licensed center began operating	Categorical (Pre-1988, 1988-2013, or Post 2014)	Pre-1988: 423 1988-2013: 2,548 After 2014: 1,032
Demographic information				
Percent free/reduced lunch	PER_FREE	Percent of enrolled children receiving free/reduced lunch	Continuous	Mean: 55% Median: 66% Max: 100% Missing: 821 (20.5%)
Percent non-White	PER_NON_WHITE	Percent of enrolled children identifying as non-White	Continuous	Mean: 53% Median: 53% Max: 100% Missing: 1,108 (27.7%)
Total enrolled	TOTAL_ENROLL	Total number of children enrolled	Discrete	Mean: 5 Median: 4 Max: 65 Missing: 562 (14%)
Head Start	head_start	Whether the center was a federally funded Head Start program	Binary	Head Start: 521 Not Head Start: 3,482
Opportunity Zone	opzone	Whether the center was located in an area designated as a HUD Opportunity Zone	Binary	In Opportunity Zone: 570 Not in Opportunity Zone: 3,434

Table S2. Additional variables included in the data set compiled from publicly available data sources.

Variable Name	Abbreviated name	Description	Variable type	Summary	Data source
Past LCR exceedance	any_lcr_exceedance	Whether the facility was served by a water system with an action level exceedance (≥10% of collected water samples collected by the utility exceeding 15 ppb) in the five years prior to Clean Water for Carolina Kids sampling (2016-2021). If served by a private well, coded as "LCR NA" but not missing.	Categorical	Past exceedance: 55 No past exceedance: 2,889 LCR NA: 4.3% Missing: 884 (22%)	EPA Safe Drinking Water Information System[16]
Source water type	type_binary	Water system source water type. Groundwater use indicates groundwater alone; if a water system used any surface water it was coded as surface water. Centers reporting using private well water default to groundwater. Other centers not served by a community water system coded as missing.	Binary	Groundwater: 610 Surface water: 2,772 Missing: 622 (15.5%)	North Carolina Drinking Water Watch[17]
# connections of water system	connections_cat	Number of service connections in the water system. Centers reporting using private well water default to 1 service connection. Other centers not served by a community water system coded as missing.	Categorical	<3300 (Small, very small, and private systems): 766 3301-10000 (Medium systems): 744 >10001 (Large, very large systems): 1971 Missing: 622 (15.5%)	North Carolina Drinking Water Watch[17]
Phosphate addition	Phos_binary	Whether the water system implements phosphate-based corrosion control. Centers reporting using private well water default to no phosphate addition. Other centers not served by a community water	Binary	Phosphate addition: 1,944 No phosphate addition: 1,438 Missing: 622 (15.5%)	North Carolina Drinking Water Watch[17]

		system coded as missing.			
pH adjustment	pH_binary	Whether the water system implements pH adjustment. Centers reporting using private well water default to no pH adjustment. Other centers not served by a community water system coded as missing.	Binary	pH adjustment: 2,481 No pH adjustment: 901 Missing: 622 (15.5%)	North Carolina Drinking Water Watch[17]
Coagulation	coagulation	Whether the water system implements coagulation in the treatment train. Centers reporting using private well water default to no coagulation. Other centers not served by a community water system coded as missing.	Binary	Coagulation: 1,947 No coagulation: 1,435 Missing: 622 (15.5%)	North Carolina Drinking Water Watch[17]
Chloramines	chloramines	Whether the water system uses chloramination for disinfection. Centers reporting using private well water default to no chloramination. Other centers not served by a community water system coded as missing.	Binary	Chloramines: 1,032 No chloramines: 2,350 Missing: 622 (15.5%)	North Carolina Drinking Water Watch[17]
Urbanicity	ruca_cat	Whether the center was located in a metropolitan, micropolitan, small town, or rural area based on Rural-Urban Commuting Area (RUCA) code	Categorical	Metropolitan: 2,920 Micropolitan: 669 Rural: 161 Small town: 24 Missing: 8 (0.1%)	US Department of Agriculture, Economic Research Service[18]
Block group median household income	med_hh_income_cbg	Median household income of the block group where each facility was located	Continuous	Mean: 56,031 Median: 49,458 Max: 250,001 Missing: 174 (4.3%)	American Community Survey 2020[19]
Block group educational attainment	perc_hs_higher_cbg	Proportion of block group population having attained a high school degree or higher	Continuous	Mean: 0.13 Median: 0.11 Max: 0.77 Missing: 11 (0.2%)	American Community Survey 2020[19]
Block group % non-White	perc_non_white_cbg	Proportion of block group population that identified as a race	Continuous	Mean: 0.39 Median: 0.34 Max: 1.0	American Community

		category other than White alone		Missing: 11 (0.2%)	Survey 2020[19]

Table S3. Summary of all performance metrics for each model.

Model name	Prior probability of model target	AU-ROC full training set	Mean AU-ROC 10-fold cross validation on training set	AU-ROC test set	AU-PR training set	Mean AU-PR 10-fold cross validation on training set	AU-PR test set	Max F1-score	Max F2-score
Max>1	0.56	0.75	0.73	0.70	0.78	0.76	0.73	0.76	0.87
Max>5	0.21	0.72	0.71	0.76	0.43	0.43	0.38	0.48	0.62
Max>10	0.13	0.75	0.72	0.74	0.32	0.31	0.26	0.38	0.52
Max>15	0.09	0.73	0.69	0.76	0.24	0.24	0.20	0.31	0.43
P90>1	0.49	0.71	0.68	0.66	0.70	0.68	0.63	0.69	0.84
P90>5	0.15	0.71	0.67	0.71	0.31	0.31	0.22	0.37	0.52
P90>10	0.07	0.72	0.65	0.75	0.21	0.20	0.14	0.25	0.37
P90>15	0.05	0.67	0.62	0.69	0.12	0.16	0.10	0.22	0.28

Table S4. Summary of mean F1 and F2 scores for each heuristic from 10-fold cross validation.

	F1 scores			F2 scores		
Heuristic	**Mean**	**Range**	**BN model mean % improvement**	**Mean**	**Range**	**BN model mean % improvement**
Groundwater	0.23	0.15-0.31	80%	0.23	0.20-0.25	142%
Head Start	0.23	0.14-0.28	83%	0.22	0.19-0.25	155%
Pre-1988 buildings	0.28	0.11-0.53	65%	0.36	0.22-0.53	51%
Private well water	0.14	0.10-0.19	248%	0.11	0.07-0.16	506%

Table S5. Summary table of all significant variables included in each model. Interactive model structures can also be seen at www.cleanwaterforcarolinakids.org/publications/bn_models.

Model	Variables included	Model	Variables included
Max>1	% free/reduced lunch enrollment	**P90>1**	% free/reduced lunch enrollment
	% non-White enrollment		# samples
	Total enrollment		% taps filtered
	# samples		Head Start
	% taps filtered		School-based
	Head Start		Home-based
	School-based		Past faucet fixture change
	Home-based		Year of past faucet fixture change
	Past faucet fixture change		Year center began operating
	Year of past faucet fixture change		Source water type
	Year center began operating		pH adjustment
	Source water type		Chloramination
	pH adjustment		# connections of water system
	Chloramination		Urbanicity
	# connections of water system		Block group median household income
	Urbanicity	**P90>5**	% free/reduced lunch enrollment
Max>5	% free/reduced lunch enrollment		# samples
	Total enrollment		Head Start
	# samples		School-based
	Head Start		Home-based
	School-based		Past faucet fixture change
	Home-based		Source water type
	Past faucet fixture change		# connections of water system
	Year center began operating		Past LCR exceedance
	Source water type		Block group % non-White
	# connections of water system		On-site wastewater system
Max>10	% free/reduced lunch enrollment	**P90>10**	% free/reduced lunch enrollment
	# samples		# samples
	Head Start		Head Start
	School-based		School-based
	Home-based		Home-based
	Past faucet fixture change		Past faucet fixture change
	Source water type		Source water type
	# connections of water system		pH adjustment
	Past LCR exceedance		Past LCR exceedance
	Phosphate addition		Phosphate addition
Max>15	% free/reduced lunch enrollment		Block group % non-White
	# samples		On-site wastewater system

	Head Start		Block group educational attainment
	School-based	P90>15	% free/reduced lunch enrollment
	Home-based		# samples
	Past faucet fixture change		Head Start
	Year center began operating		Source water type
	Source water type		pH adjustment
	pH adjustment		Past LCR exceedance
	Past LCR exceedance		On-site wastewater system
	Phosphate addition		
	Block group % non-White		

References

(1) Gonzalez-Abril, L.; Cuberos, F. J.; Velasco, F.; Ortega, J. A. Ameva: An Autonomous Discretization Algorithm. *Expert Syst Appl* **2009**, *36* (3 PART 1), 5327–5332. https://doi.org/10.1016/j.eswa.2008.06.063.

(2) Ropero, R. F.; Renooij, S.; van der Gaag, L. C. Discretizing Environmental Data for Learning Bayesian-Network Classifiers. *Ecol Modell* **2018**, *368*, 391–403. https://doi.org/10.1016/j.ecolmodel.2017.12.015.

(3) García, S.; Luengo, J.; Sáez, J. A.; López, V.; Herrera, F. A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning. *IEEE Trans Knowl Data Eng* **2013**, *25* (4), 734–750. https://doi.org/10.1109/TKDE.2012.35.

(4) Conrady, S.; Jouffe, L. *Bayesian Networks and BayesiaLab: A Practical Introduction for Researchers*; Bayesia USA: Franklin, TN, 2015.

(5) Roostaei, J.; Colley, S.; Mulhern, R.; May, A. A.; Gibson, J. M. Predicting the Risk of GenX Contamination in Private Well Water Using a Machine-Learned Bayesian Network Model. *J Hazard Mater* **2021**, *411* (October 2020), 125075. https://doi.org/10.1016/j.jhazmat.2021.125075.

(6) Mulhern, R.; Roostaei, J.; Schwetschenau, S.; Pruthi, T.; Campbell, C.; MacDonald Gibson, J. A New Approach to a Legacy Concern: Evaluating Machine-Learned Bayesian Networks to Predict Childhood Lead Exposure Risk from Community Water Systems. *Environ Res* **2022**, *204*, 112146. https://doi.org/10.1016/j.envres.2021.112146.

(7) Dong, Y.; Peng, C. Y. J. Principled Missing Data Methods for Researchers. *Springerplus* **2013**, *2* (222), 1–17. https://doi.org/10.1186/2193-1801-2-222.

(8) Kuhns, C.; Martinchek, K.; Gupta, P. *Combating Food Insecurity and Supporting Child Nutrition through the Child Care Sector*; Washington D.C., 2021. https://www.urban.org/sites/default/files/publication/105106/combating-food-insecurity-and-supporting-child-nutrition-through-the-child-care-sector.pdf.

(9) Madley-Dowd, P.; Hughes, R.; Tilling, K.; Heron, J. The Proportion of Missing Data Should Not Be Used to Guide Decisions on Multiple Imputation. *J Clin Epidemiol* **2019**, *110*, 63–73. https://doi.org/10.1016/j.jclinepi.2019.02.016.

(10) van Buuren, S.; Groothuis-Oudshoorn, K. Mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw* **2011**, *45* (3), 1–67. https://doi.org/10.18637/jss.v045.i03.

(11) Ke, X.; Keenan, K.; Smith, V. A. Treatment of Missing Data in Bayesian Network Structure Learning: An Application to Linked Biomedical and Social Survey Data. *BMC Med Res Methodol* **2022**, *9*, 1–2. https://doi.org/10.1186/s12874-022-01781-9.

(12) Scutari, M. *Structure learning from missing data*. bnlearn - an R package for Bayesian network learning and inference. https://www.bnlearn.com/documentation/man/structural.em.html (accessed 2023-02-09).

(13) Scutari, M. *Parameter learning from data with missing values*. bnlearn - an R package for Bayesian network learning and inference. https://www.bnlearn.com/examples/missing-fitting/ (accessed 2023-02-09).

(14) Scutari, M. *Predict or impute missing data from a Bayesian network*. bnlearn - an R package for Bayesian network learning and inference. https://www.bnlearn.com/documentation/man/predict.and.impute.html (accessed 2023-02-09).

(15) Hoponick Redmon, J.; Kondash, A. J.; Norman, E.; Johnson, J.; Levine, K.; McWilliams, A.; Napier, M.; Weber, F.; Stella, L.; Wood, E.; Lee Pow Jackson, C.; Mulhern, R. Lead Levels in Tap Water at Licensed North Carolina Child Care Facilities, 2020-2021. *Am J Public Health* **2022**, *112* (S7), S695–S705. https://doi.org/10.2105/AJPH.2022.307003.

(16) EPA. *Safe Drinking Water Information System*. https://sdwis.epa.gov/ords/sfdw_pub/f?p=108:200 (accessed 2023-02-07).

(17) NCDEQ. *Drinking Water Watch*. https://www.pwss.enr.state.nc.us/NCDWW2/ (accessed 2020-07-05).

(18) USDA Economic Research Service. *Rural-Urban Commuting Area Codes*. https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes.aspx (accessed 2022-08-15).

(19) Manson, S.; Schroder, J.; van Riper, D.; Kugler, T.; Ruggles, S. *IPUMS National Historical Geographic Information System: Version 17.0*; Minneapolis, MN, 2022. https://doi.org/10.18128/D050.V17.0.