Environmental Research 204 (2022) 112146

Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

A new approach to a legacy concern: Evaluating machine-learned Bayesian networks to predict childhood lead exposure risk from community water systems

Riley Mulhern [a,1,*], Javad Roostaei [a,2], Sara Schwetschenau [b,3], Tejas Pruthi [a], Chris Campbell [c], Jacqueline MacDonald Gibson [d]

[a] Department of Environmental Sciences and Engineering, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, 135 Dauer Drive, Chapel Hill, NC, 27599, USA
[b] Department of Civil and Environmental Engineering, College of Engineering, Wayne State University, 5050 Anthony Wayne Dr., Detroit, Michigan, 48202, USA
[c] Environmental Working Group, 1436 U St. NW, Suite 100, Washington, DC, 20009, USA
[d] Department of Environmental and Occupational Health, School of Public Health, Indiana University, 1025 East 7th Street, Bloomington, IN, 47405, USA

ARTICLE INFO

Keywords:
Bayesian networks
Drinking water
Blood lead levels
Machine learning
Risk assessment
Health disparity

ABSTRACT

Lead in drinking water continues to put children at risk of irreversible neurological impairment. Understanding drinking water system characteristics that influence blood lead levels is needed to prevent ongoing exposures. This study sought to assess the relationship between children's blood lead levels and drinking water system characteristics using machine-learned Bayesian networks. Blood lead records from 2003 to 2017 for 40,742 children in Wake County, North Carolina were matched with the characteristics of 178 community water systems and sociodemographic characteristics of each child's neighborhood. Bayesian networks were machine-learned to evaluate the drinking water variables associated with blood lead levels ≥2 μg/dL and ≥5 μg/dL. The model was used to predict geographic areas and water utilities with increased lead exposure risk. Drinking water characteristics were not significantly associated with children's blood lead levels ≥5 μg/dL but were important predictors of blood lead levels ≥2 μg/dL. Whether 10% of water samples exceeded 2 ppb of lead in the most recent year prior to the blood test was the most important water system predictor and increased the risk of blood lead levels ≥2 μg/dL by 42%. The model achieved an area under the receiver operating characteristic curve of 0.792 (±0.8%) during ten-fold cross validation, indicating good predictive performance. Water system characteristics may thus be used to predict areas that are at risk of higher blood lead levels. Current drinking water regulatory thresholds for lead may be insufficient to detect the levels in drinking water associated with children's blood lead levels.

1. Introduction

Lead has been used to deliver piped drinking water for millennia due to its unique chemical properties, including low melting point, malleability, and relative resistance to corrosion. Despite warnings about the negative health effects of lead exposure since antiquity (Hodge, 1981; Lessler, 1988; Vuorinen et al., 2019), by the 19th century, 70 percent of drinking water mains and service lines in the United States (U.S.) contained lead (Rabin, 2008). The effects of increased blood lead levels in infants and young children include neurological damage resulting in permanent developmental, learning, and IQ deficits (Bellinger et al., 1987; Canfield et al., 2003; Lanphear et al., 2005; McMichael et al., 1988; Needleman et al., 1990). Higher blood lead concentrations have also been significantly associated with preeclampsia (Poropat et al., 2018) and excess mortality among adults (Lanphear et al., 2018). Despite documented cases of lead poisoning in households served by

* Corresponding author.
E-mail address: rmulhern@rti.org (R. Mulhern).
[1] Present address: RTI International, Center for Environmental Health, Risk, and Sustainability, 3040 East Cornwallis Road, P.O. Box 12194, Research Triangle Park, NC 27709.
[2] Present address: Hazen & Sawyer, 4011 WestChase Boulevard, Suite 500, Raleigh, NC 27607, USA.
[3] Present address: Columbia Water Center, Columbia University, 500 West 120th St New York, New York, 10027, USA

https://doi.org/10.1016/j.envres.2021.112146
Received 29 June 2021; Received in revised form 24 September 2021; Accepted 26 September 2021
Available online 29 September 2021

lead pipes as early as the 1850s (Adams, 1859), plumbers and pipe manufacturers continued to promote the use of lead in U.S. drinking water systems well into the 20th century (Gray, 1916; Rabin, 2008).

Lead plumbing was first regulated in the U.S. in 1986 in an amendment to the Safe Drinking Water Act which required components used for drinking water to contain no more than 8% lead, whereas, previously, fittings and fixtures may have contained 40–50% lead (Maas et al., 2005). The subsequent Lead and Copper Rule, promulgated in 1991 and revised in 2000, 2004, and 2007 requires water sampling for lead at a small number of selected individual residences and establishes an action level of 15 parts per billion (ppb) (USEPA, 1991, 2000, 2004, 2007). Under the Lead and Copper Rule action must be taken if more than 10% of households tested in a designated year (i.e., the 90th percentile of monitoring samples) exceed the action level. If this threshold is exceeded, the utility is required to implement system-wide corrosion control practices and disseminate educational materials to the public. If treatment is inadequate, the water system may be required to remove lead service lines. In 2019, revisions to the Lead and Copper Rule were proposed that would establish an additional trigger level at 10 ppb. If 10% of samples exceed the trigger level, utilities would be required to perform a corrosion control optimization study as a preemptive measure to identify a strategy for lead mitigation (USEPA, 2019).

Although the Lead and Copper Rule acknowledges that there is no safe level of lead in drinking water, the regulation is principally designed to monitor utility corrosion control practices, rather than to be protective of health at the household scale. Corrosion control treatment is assessed through sample collection at a limited number of residences, up to 100 locations for systems serving greater than 100,000 people (i.e., 0.1% of homes served). Systems in compliance with the Lead and Copper Rule are not required to remove lead service lines. As a result, estimates suggest that between 6.1 and 12.8 million lead service lines still exist in approximately 30% of all U.S. community water systems, serving between 15 and 21 million people (Cornwell et al., 2016; NRDC, 2021). What is more, as many as 77% of all U.S. housing units (over 80 million homes) likely contain lead solder joints and virtually all U.S. homes contain some brass components with up to 8% lead by weight (Triantafyllidou and Edwards, 2012). Thus, even where utilities are in compliance with all Lead and Copper Rule provisions, the conditions influencing lead release within distribution systems are often poorly characterized (Schwetschenau et al., 2020). Legacy lead service lines, lead-bearing plumbing components, and low sampling rates can all cause unsafe lead levels at individual household taps to remain undetected (Riblet et al., 2019; Triantafyllidou and Edwards, 2012). Consequently, compliance with the Lead and Copper Rule action level is not considered adequately protective for children and formula-fed infants, who represent the population most vulnerable to lead exposure (Lambrinidou et al., 2010; Redmon et al., 2018; Triantafyllidou and Edwards, 2012).

Despite the limitations of the Lead and Copper Rule, it remains the only regulatory tool available to manage drinking water lead exposure from the myriad potential sources of lead in drinking water distribution and premise plumbing systems. Therefore, it is critical to develop improved risk assessment techniques that can be implemented in conjunction with the Lead and Copper Rule framework and that are based in an understanding of how system-wide conditions bear upon public health outcomes. Detailed mechanistic models relating drinking water to children's blood lead levels are not always practical for population-wide predictions given that household drinking water lead concentrations are often highly variable and difficult to model accurately (Del Toral et al., 2013; Trueman et al., 2016). Thus, new approaches are needed for system-level analysis and risk assessment that go beyond Lead and Copper Rule compliance and proactively identify and respond to lead exposure risk in community water systems where it may otherwise be overlooked.

Machine learning approaches provide a promising alternative to mechanistic models due to the ability to leverage diverse datasets to predict increased lead exposure risk without expensive sampling or comprehensive household level data. In this work, we tested machine-learned Bayesian network models as one possible approach that provides several advantages over traditional statistical techniques. First, by using a nonparametric modeling approach, Bayesian networks can capture complex and nonlinear relationships and avoid problems of multicollinearity (Lee et al., 2019; Sebastiani and Perls, 2008). Additionally, by exploiting conditional independencies between variables in the network, Bayesian networks reduce the number of parameters required to express the joint probability distribution and thus provide a compact means of describing complex data sets (Goldszmidt, 2011).

Previous applications of Bayesian networks to assess environmental health risks include the temporal spread of West Nile Virus (Orme-zavaleta et al., 2006), exposure to *Staphylococcus aureus* in pasteurized milk (Barker, 2013), low birth weight from arsenic exposure in drinking water (Zabinski et al., 2016), and transmission of Ebola virus through municipal wastewater systems (Zabinski et al., 2018). Francis et al. have also demonstrated the benefits of Bayesian networks in predicting pipe breaks in drinking water distribution systems (Francis et al., 2014). Potash et al. present a promising application of other machine learning approaches for predicting risk of lead exposure from household paint and dust in Chicago (Potash et al., 2015), but similar tools are needed for drinking water. Recent advancements in the use of machine learning and Bayesian networks for evaluating lead exposure risk in drinking water include predicting water lead contamination in schools in California and Massachusetts (Lobo et al., 2021) and in private wells in Virginia (Fasaee et al., 2021), but, to our knowledge, this is the first study to apply these methods to community water systems using paired drinking water and blood lead data.

Thus, in this study, we tested the use of machine-learned Bayesian networks to model the complex probabilistic relationships between system-level drinking water characteristics, Lead and Copper Rule monitoring data, and children's blood lead levels. Our specific objectives were to:

1. Evaluate the performance of machine-learned Bayesian network models to predict blood lead levels from water system characteristics.
2. Assess the extent to which children's blood lead levels are associated with community water system characteristics.

This approach will help drinking water and public health managers identify and predict increased blood lead risk in community water system service areas and prioritize interventions where they are most likely to be needed.

2. Methods

2.1. Summary

Bayesian networks were leveraged in conjunction with geospatial analysis to identify the relationships between water system characteristics, geographic and demographic characteristics and blood lead surveillance data in Wake County, North Carolina. Our final data set for machine learning included 1) blood lead levels from 40,742 children served by community water systems between 2003 and 2017; 2) neighborhood demographic and socioeconomic data for each child's address; 3) Lead and Copper Rule monitoring data between 2003 and 2017 from 178 community water systems in the county; and 4) characteristics of each water system including size, infrastructure, and treatment train. Each child was paired with the characteristics of the water utility he or she was served by in order to identify relationships between water system operational parameters, water lead concentrations, and children's blood lead levels. Fig. 1 summarizes the spatial relationships between these integrated data sets. Fig. 2 summarizes the

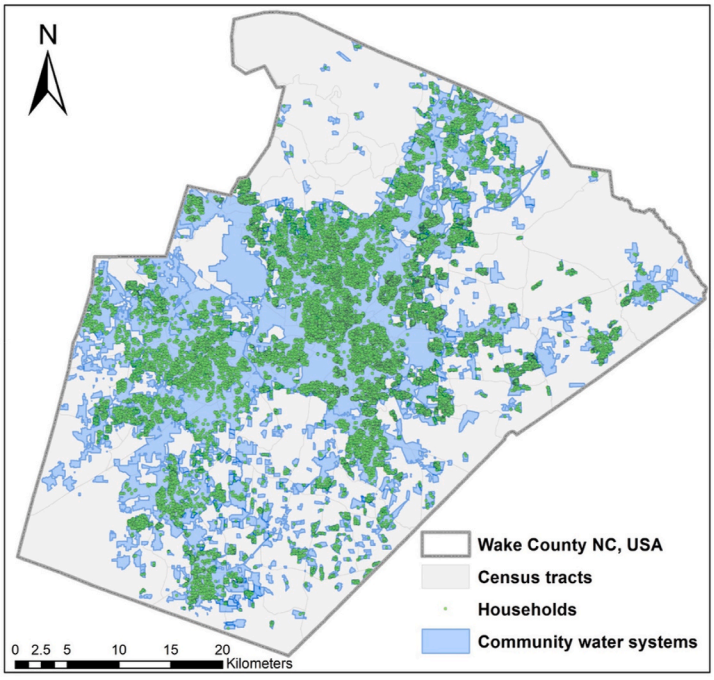

Fig. 1. Spatial relationship among integrated datasets. Demographic and socioeconomic characteristics are described at the Census block group level. For visual clarity, only Census tracts are shown on the map with multiple block groups contained within a Census tract. Census tracts are outlined in grey. Community water system service areas are shown in blue. Individual households of children with blood lead tests in the NC Childhood Lead Poisoning Prevention data set are shown as green dots. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

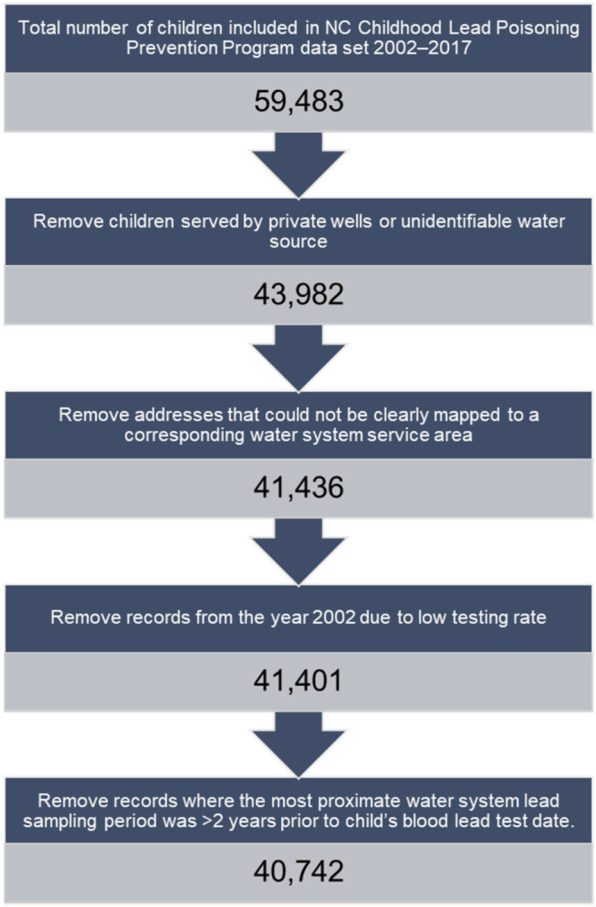

Fig. 2. Steps taken to refine the North Carolina Lead Poisoning Prevention Program data set of childhood blood lead tests for machine learning.

steps take to refine the blood lead data set for machine learning. The data sources and the integration procedure used are described briefly below, with more detailed information available in the Supplemental Information (SI), Section S1.

2.2. Data set integration and compilation

2.2.1. Blood lead levels, household attributes, and neighborhood characteristics

Children's blood lead measurements from 2002 to 2017 were obtained for Wake County, NC from the NC Department of Health and Human Services Childhood Lead Poisoning Prevention Program. These data represent 59,483 blood lead test results including the household address, birth date, and gender of each child tested. Child ages ranged 0–73 months. From 2005 to 2017, the NC Childhood Lead Poisoning Prevention Program screened approximately one half of all eligible North Carolina Children (Angelon-Gaetz and Newman Chelminski, 2018; NCDHHS, 2021). Although children's blood lead levels have also been shown to exhibit seasonal variation, with higher levels tending to occur in the summer months (Haley and Talbot, 2004; Yiin et al., 2000), the influence of season could not be assessed as only the year of the blood lead test was provided for the machine learning data set. The use of these data was approved by the University of North Carolina Institutional Review Board.

Wake County residential property tax records were used to identify the value of each child's home and the water source used (i.e., private well or community system). Children served by private wells, approximately 15%, were removed (Fig. 2). Each household was also matched to the demographic and socioeconomic characteristics for the corresponding U.S. Census block group from the American Community Survey (2013–2017). The georeferencing procedure and validation method for this data set has been described in detail elsewhere (Macdonald Gibson et al., 2020).

In order to develop the machine learned model to predict lead risk (see section 2.3), recorded blood lead levels were classified by whether they were greater than or equal to two thresholds: 5 μg/dL and 2 μg/dL. The 5 μg/dL target was selected as it is the Centers for Disease Control and Prevention (CDC) Reference Level for clinically determining elevated blood lead, while the 2 μg/dL threshold was chosen as the

median blood lead level in the data set (including children receiving water from private wells) in order to also detect possible relationships with subclinical lead exposures. Subclinical effects are significant as they may not result in obvious clinical symptoms but may still result in negative neurological outcomes not observed until later during a child's development (National Toxicology Program, 2012). The trend in children's blood lead levels served by community water systems is shown in Fig. 3, Panel A. As can be seen, there is a decreasing trend in the proportion of children with blood lead levels $\geq$5 μg/dL and $\geq$ 2 μg/dL over time (records from 2002 were removed due to low testing rate in that year).

2.2.2. Water lead levels and water system characteristics

From the original NC Childhood Lead Poisoning Prevention Program data set containing 59,483 records, 41,401 children could be matched to 178 water systems in Wake County after filtering out children served by private wells, children whose water source could not be identified, addresses that could not be unambiguously mapped to a water system, and records from 2002 (Fig. 2). The 178 water systems range in size from very small groundwater systems serving unincorporated subdivisions with fewer than 100 connections to large surface water utilities. While most children (85%) were served by two large surface water utilities, 96% of the water systems in the county have a groundwater source, and 97% are considered small or very small systems. Additional water lead level summary statistics by water system are shown in Figure S1 and Table S2. The approximate service area of each water utility was determined through publicly available maps of subdivisions and city limits in Wake County. For community water systems associated with incorporated cities, the full extent of the city limits was used as the approximate utility service area. Community water systems associated with unincorporated subdivisions were determined to serve the subdivision boundaries that share the community water system's name. A flow chart outlining the decision-making logic for determining the service area of each water system serving unincorporated subdivisions is provided in Figure S2. Each child was then mapped to their providing water system by identifying the service area containing each child's address. The approximate service areas of all mapped water systems in the county are shown in Figure S3.

Characteristics describing each community water system were obtained from public records through the NC Department of Environmental Quality public water supply system registry (NCDEQ, 2020) which includes source water type (e.g., groundwater, surface water, or purchased water), number of service connections, age of the system, system infrastructure (e.g., the number of storage tanks, wells, and surface water intakes), and relevant treatment processes (e.g., pH adjustment, phosphate-based corrosion control, and disinfectant type). In addition, the Environmental Working Group's Tap Water Database was used to access the Lead and Copper Rule water quality monitoring data for each system, resulting in 13,664 individual sample results from 2002 to 2017 (EWG, 2019). Samples taken after 2009 were specifically identified as being from household taps within the distribution system. Pre-2009 samples were assumed to also be from the distribution system, although these samples lacked this coding in the database.

As can be seen in Fig. 3, Panel B, the water lead levels among the 178 systems analyzed fluctuated over time. Because many water systems in the data set qualified for reduced (i.e., triennial) sampling under the

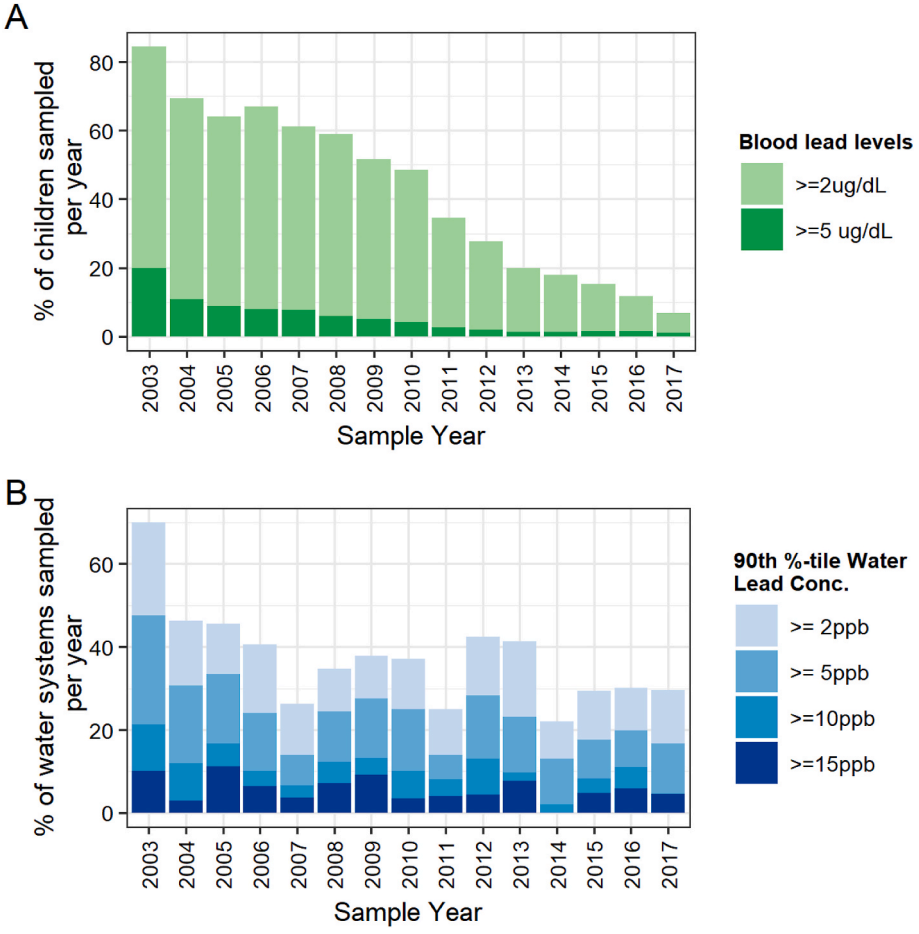

Fig. 3. Summary of blood and water lead concentrations in Wake County, NC from 2003 to 2017. Panel A: Percent of children that are served by community water systems with blood lead levels greater than or equal to 2 μg/dL and 5 μg/dL each year. Panel B: Percent of water systems serving these children for which the 90th percentile water sampling result is greater than the specified thresholds (2 ppb, 5 ppb, 10 ppb, and 15 ppb).

Lead and Copper Rule, the specific water systems sampled in each year varied. This explains some of the inter-year variation in 90th percentile lead concentrations seen in Fig. 3, Panel B. For each year of water samples collected by each water system, the mean, median, standard deviation, and the percentage of samples exceeding selected thresholds were calculated and assumed to be representative of interim years when water sampling was not conducted for that system. This assumption allowed for a balanced panel dataset to be created. The selected thresholds included the Lead and Copper Rule action level (15 ppb), the proposed trigger level (10 ppb), and three additional thresholds: 1 ppb (the lowest reporting limit used in the region), 2 ppb, and 5 ppb. Values below the reporting limit for a specific water system were reported as non-detects. In the absence of detailed information regarding the reporting limit used by each water utility, which ranges from 1 to 5 ppb, non-detects were substituted with a value of zero (see SI Section 1 for additional detail). This substitution was considered a conservative assumption since the potential effect would be to bias water lead levels low, so the risk of a Type I error (i.e., if a significant effect of water lead levels on blood lead levels were observed where it did not exist) was low. These summary statistics and the assumption that they would be representative of water quality in interim years between sampling periods allowed system-level water lead level data to be linked to individual blood lead level results by year.

2.2.3. Final machine learning data set

Since water lead sampling did not occur in every year blood lead measurements were available, data from the most proximate sample year prior to each child's blood lead test were joined to the blood lead result. For example, the town of Cary performed Lead and Copper Rule monitoring every three years from 2003 to 2015 (Table S2). A child in Cary who had his or her blood tested for lead in 2010 was thus paired with the water lead data from 2009 as the most recent prior system-wide sampling date. The data set of 41,401 children matched to water systems was then restricted further to consider only children for whom the most proximate water lead sampling period was within the two years prior of the child's blood lead test, resulting in a final data set of 40,742 unique records (Fig. 2). Two years was selected as a reasonable time frame that retained the majority of records while eliminating outliers for whom no prior or recent water lead concentration data were available (see Section S1). Although the half-life of lead in blood has been shown to be only 1–2 months, elevated blood lead levels in children may require over a year to decline (Dignam et al., 2008), and a second half-life of lead in blood due to the replenishment of lead stored in the bones can be up to four years (ATSDR, 2007). Further, high water lead levels may go undetected and treatment techniques to reduce water lead exposures may not be adjusted annually due to reduced sampling, thus causing water lead exposures to be prolonged for children served by some systems.

The final data set included 60 matched household, demographic, socioeconomic, and water system variables for each child. The full list of variables can be found in Table S1. Summary statistics for each variable are provided in Tables S3, S4 and S5. The data transformation process resulted in a flattened database for machine learning that was propositionalized from the multiple relational data sets described in the previous sections (Kramer et al., 2001). As discussed by Maier et al. (2013), propositionalization of relational data (i.e., data that violates the assumptions of independence and identical distribution (IID)) to define variables prior to machine learning has limitations for causal inference, but can generally be used for the purposes of predictive inference and evaluating statistical associations in complex, relational systems with non-IID data without loss of predictive accuracy.

2.3. Bayesian network theory and model construction

Bayesian networks are probabilistic graphical models in the form of directed acyclic graphs that allow complex joint probability distributions to be graphically represented. The qualitative component of Bayesian networks includes a set of random variables, $\{V_1, \ldots, V_n\}$, each represented graphically as a node and connected by arcs indicating statistical dependencies among variables. The quantitative component behind the graphical representation is the joint probability distribution over V. Bayesian networks enable complex joint probability distributions to be decomposed into the product of the conditional distribution of each V_i given its "parent" nodes in the graph (the nodes with arrows pointing directly to V_i), represented as $Pa(V_i)$, such that:

$$P(V) = \prod_{ieV} P(V_i | Pa(V_i))$$

In this study, Bayesian networks were constructed and evaluated using the software *BayesiaLab* (Changé, France). Upon importing the database to the software, continuous variables were discretized using a built-in discretization algorithm (R2-GenOpt) that maximizes the variance of the discretized variable explained by its corresponding continuous variable (Bayesia, 2021). The data set had no missing values for blood lead levels or water system characteristics, but some missingness at random for the variables child gender (2.1%), home value (5.7%), and year of home construction (6.9%) (Table S3). Missing values were inferred using a structural expectation-maximization algorithm which uses dynamic imputation with weighted observations to infer missing values based on the structure of the network (Conrady and Jouffe, 2015; Friedman, 1997). To train the network, we used a series of supervised learning algorithms to assess the probability of a child's blood lead level meeting or exceeding 5 μg/dL or 2 μg/dL (referred to as the "target"). Based on this probability, the model then classified each child as above or below each threshold. Classification models used to distinguish between discrete classes are central to machine learning (Kotsiantis, 2007) and are used in a wide range of problem domains, ranging from the classification of e-mail as Spam (Guzella and Caminhas, 2009) to medical diagnostics, including heart failure (Olsen et al., 2020), breast cancer (Hu et al., 2020), and COVID-19 (Li et al., 2020). The use of classification models in this study provides similarly useful decision-making information regarding lead exposure.

BayesiaLab includes a variety of built-in machine learning algorithms including Naïve Bayes, augmented Naïve Bayes, tree augmented Naïve Bayes, Markov blanket, and augmented Markov blanket. These algorithms use a greedy search strategy to test linkages between nodes that reduce complexity while maximizing predictive capability as measured by the minimum description length score (Conrady and Jouffe, 2015). The significance of linkages is evaluated in the software using the G-test statistic (McDonald, 2014). The structural coefficient $\boldsymbol{\alpha}$ was adjusted in *BayesiaLab* to 0.35 based on visual inspection of the structure/target precision ratio (Conrady and Jouffe, 2015). Each algorithm was tested separately in *BayesiaLab* and the highest performing model was selected. Performance was evaluated using ten-fold cross-validation to assess the area under the receiver operating characteristic (ROC) curve, which plots the relationship between the true positive rate and false positive rate for different probability thresholds classifying the binary outcome of interest (in this case, whether a child's blood lead level will meet or exceed each target). An ROC score of 1 indicates a model that perfectly discriminates between each possible outcome, while an ROC score of 0.50 indicates a useless model, where the predictive ability is no better than chance (Carter et al., 2016). The model with the highest ROC score was selected as the final model.

The total effect of each predictor variable was then assessed by comparing the prior probability (i.e., the marginal or unconditional probability) of a child's blood lead level exceeding each target to the posterior probability (i.e., the conditional probability of the target given the values of each predictor node). When calculating the posterior probability, information is allowed to flow freely through all connected nodes. Therefore, assessing this effect does not assume causality and only measures the strength of the association given the other variables in the network. The difference in the uncertainty between the prior and posterior states of the target given each predictor is known as the mutual

information and provides an additional measure of which variables have the greatest predictive importance (Conrady and Jouffe, 2015).

3. Results

3.1. Association between blood lead levels and drinking water system characteristics

3.1.1. Elevated blood lead levels (5 μg/dL target)

The prior probability of exceeding the CDC Reference Level of 5 μg/dL in the county was 4.4%, meaning that, without additional information, any child in the county had a 4.4% chance of having a blood lead level ≥5 μg/dL from 2003 to 2017. None of the water system characteristics, including treatment, infrastructure, and water lead levels, were significantly associated with the 5 μg/dL target in any of the models tested. This result indicates that these variables do not share significant mutual information with the target and thus do not reduce the uncertainty. The highest performing model for the 5 μg/dL threshold used an augmented Naïve Bayes structure (Figure S4). The ROC curve is shown in Figure S5 and the fully specified marginal probabilities found in Table S6.

The most significant predictors of the CDC Reference Level target were blood test year, child age, median household income, home value, and the proportion of the Census block group that identified as Black. A large effect of blood test year was observed, decreasing from a 19.9% probability of exceeding 5 μg/dL in 2003 to 1.06% in 2017. This decrease is consistent with a nationwide trend of declining blood Pb levels, attributed to policies that decreased or eliminated many major lead exposure sources, including gasoline, paint, food cans, and plumbing and fixtures (Dignam et al., 2019). In NC, declining blood lead also may be partially attributed to the efforts of the North Carolina Childhood Lead Poisoning Prevention Program to perform surveillance of children's blood lead levels and to conduct remediation of households where children are found to have elevated blood Pb (Angelon-Gaetz and Newman Chelminski, 2018). This decrease among Wake County children can also be observed clearly in Fig. 3A. Children living in neighborhoods with a median household income of less than $42,000 per year exhibited a 128% increase in the probability, or risk, of reaching or exceeding the CDC Reference level compared to children in wealthier areas (median household income >$116,000). Addresses within Census block groups identifying as over 50% Black were also found to have an increase in risk compared to the county's baseline. For children in neighborhoods that were >71% Black, the risk increased by 118%. Finally, children living in homes valued less than $186,000 showed a 65% increase in risk compared to homes valued greater than $532,000. These discretizations do not necessarily indicate causal threshold effects, but the inclusion of median household income and home value as significant variables by the machine learning algorithm serves to affirm previous research showing that increased blood lead levels are often associated with socioeconomic factors that reduce a family's ability to mitigate exposures (Gleason et al., 2019; Stark et al., 1982). The significance of the proportion Black of the child's Census block group also confirms previous research that has pointed out concerning racial inequities of environmental lead exposure in children (Lanphear et al., 2002; Macdonald Gibson et al., 2020; Whitehead and Buchanan, 2019).

The learning algorithm's exclusion of any water lead concentration statistic variables indicates that elevated blood lead levels among the sample of children studied in Wake County are not significantly associated with water lead from community water systems. Our data set was weighted toward the two largest water systems in the county, however, with 85% of blood lead records associated with these two systems. To address this, we restricted the data set to exclude children served by the two largest utilities and reran the learning algorithms. In a third iteration, the data set was further restricted to evaluate only groundwater systems. In these models, blood test year remained the most important predictor and water system characteristics and Lead and Copper Rule sampling results continued to be insignificant. This suggests that our assessment was not controlled by the clustering of children within certain water systems and that blood lead levels at or above the CDC Reference Level in the county are likely attributable to additional household-level exposures such as lead paint and dust.

3.1.2. Subclinical blood lead levels (2 μg/dL target)

The prior probability of blood lead levels ≥2 μg/dL in the county was 38.7%. In contrast to the 5 μg/dL target, multiple water system characteristics became significant when predicting the probability of a child's blood lead level reaching or exceeding 2 μg/dL. Other, significant variables included the blood lead test year, household attributes, and socioeconomic and demographic characteristics of the neighborhood. The highest-performing model for evaluating the 2 μg/dL target also used an augmented Naïve Bayes structure (Fig. 4). The ROC curve for this model is provided in Figure S6. Marginal probabilities of each of the nodes in the network are provided in Table S7. The size of the effect of each node on the target's posterior probability can be seen in Fig. 5. The width of the bars in Fig. 5 represents the range of the effect (and thus the overall importance) of each predictor variable on the target node. The variables in Fig. 5 are also ranked according to the amount of mutual information each node shares with the target from greatest to least. Each of the variables included in Figs. 4 and 5 exhibited statistically significant associations with the target according to G-tests of independence except for the child's gender which was forced into the model as a control based on prior research (Macdonald Gibson et al., 2020).

As can be seen in Fig. 4, blood test year, home value, median household income, and the proportion Black of the Census block group continued to be selected as significant predictors of the 2 μg/dL target. As before, blood test year had the most mutual information with the target and was the greatest overall driver of blood lead risk in the model (Fig. 5). The risk of blood lead levels ≥2 μg/dL decreased by 92% from 2003 to 2017 in the county overall, but children living in lower income housing and neighborhoods, as well as children in majority Black neighborhoods, continued to exhibit the highest risk even in later testing years. The proportion of the Census block group that identified as Hispanic was also selected as an important demographic variable for predicting blood lead risk, such that children in neighborhoods with a Hispanic population >15% experienced a 25% increase in risk of blood lead levels ≥2 μg/dL compared to the county average. The model thus highlighted ongoing socioeconomic and racial disparities in childhood lead exposures across the county that persist even at subclinical blood lead levels.

Key water system characteristics with an effect on children's blood lead levels included the number of service connections, the age of the water system, the number of wells in groundwater systems, and whether the system practices treatment techniques that may affect corrosion within the distribution system, such as coagulation, phosphate addition, and pH adjustment. The disinfectant type (i.e., whether a utility used chloramines in the distribution system) was not identified by the algorithms tested as a significant predictor. Whether 10% of water samples collected by the utility exceeded 2 ppb of lead was also statistically significant toward predicting blood lead ≥2 μg/dL. Importantly, the algorithms used to learn the model structure did not identify the Lead and Copper Rule regulatory thresholds of 15 ppb (the action level) and 10 ppb (the trigger level) as significant predictors. While these thresholds were designed to act as overall indicators of lead exposure risk within a water system, our model suggests that these thresholds may not be sensitive enough to detect the low water lead levels that may be contribute to childhood lead exposure through drinking water.

Each of these water system characteristics had different sizes and directions of effects. The curves shown in Fig. 6 describe the change in the posterior probability of the 2 μg/dL target relative to the significant water system characteristics. For continuous variables, such as system size or the number of wells, the x-axis shows the normalized change in the mean of the predictor. For binary variables, the x-axis shows the

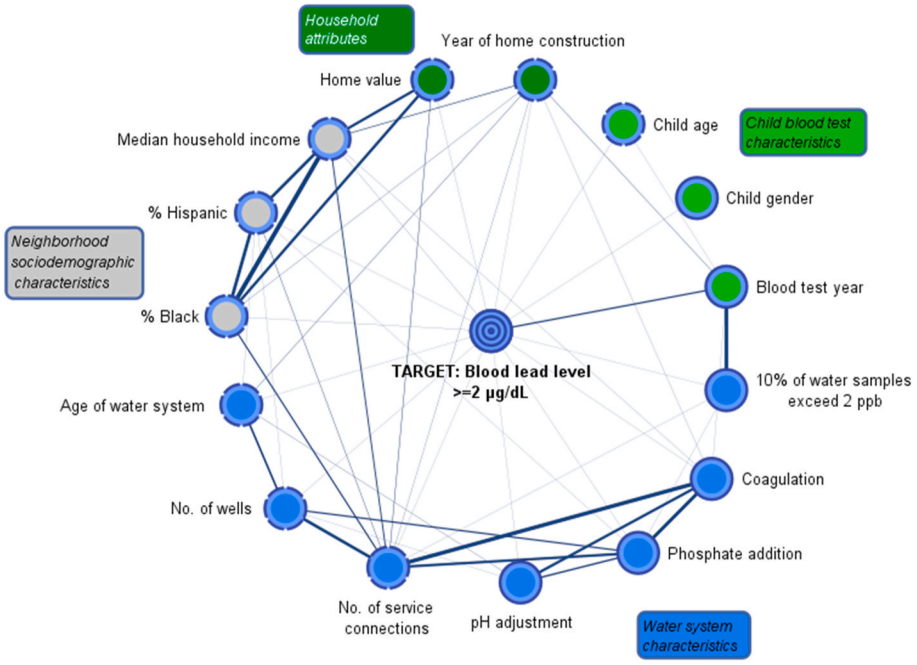

Fig. 4. Final network structure predicting the probability of each child's blood lead test result being ≥2 µg/dL. The thickness of arcs corresponds to the computed amount of mutual information between nodes. The colors of the nodes correspond to spatial scales of information shown in Fig. 1: Grey nodes represent variables at the Census block group scale; green nodes represent household and blood test characteristics at the individual address level; blue nodes represent water system characteristics at the service area scale. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

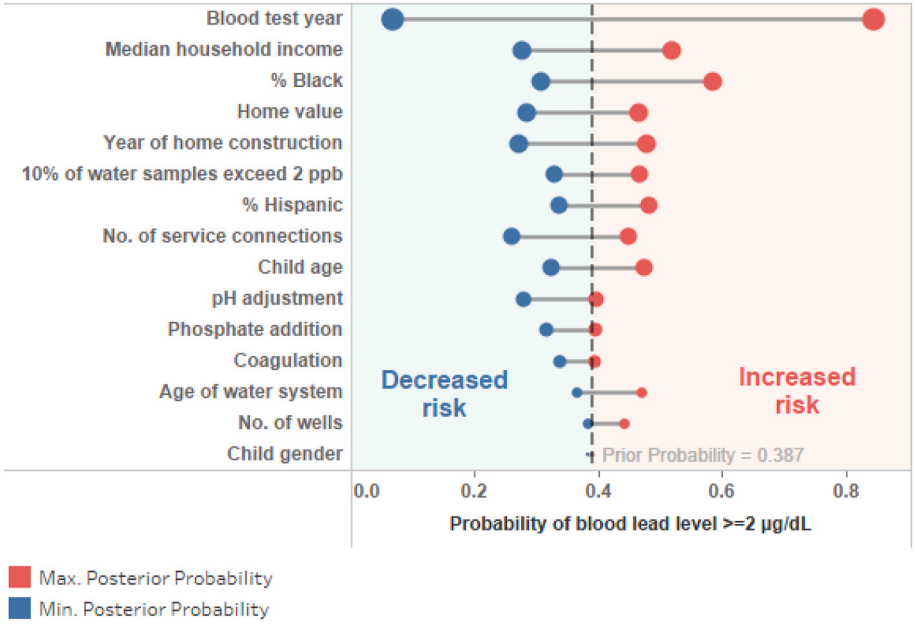

Fig. 5. Total effect of each predictor variable on the probability of blood lead level ≥2 μg/dL. Variable names are ranked from highest to lowest mutual information with the target. The size of the circles corresponds to the natural log of the mutual information for visual clarity.

proportion of children in the county served by community water systems that meet the criteria, such as using pH adjustment or exceeding 2 ppb of lead in 10% of monitoring samples. Thus, for binary variables, 100% on the x-axis indicates the risk of blood lead levels ≥2 μg/dL in the county if all children were served by systems that met the criteria, and 0% indicates the risk if none of the children were served by systems that met the criteria. The y-axis, then, gives the expected probability of blood lead levels ≥2 μg/dL with the prior probability of 38.7% shown for reference.

Most of the water system characteristics demonstrated a linear effect. For example, increasing the number of interconnected wells within a groundwater system (a measure of increasing complexity and mixing of source waters that may be treated to varying levels within the distribution system) also increased the blood lead risk. Blending of source waters has previously been shown to impact the nature of lead release within drinking water distribution systems (Tang et al., 2006). Notably, children served by utilities that do not practice phosphate-based corrosion control (n = 165 (93%) or pH adjustment (n = 70 (39%), Table S9) exhibited a decrease in the posterior probability of blood lead ≥2 μg/dL. This finding does not indicate that these treatment measures cause greater lead exposure risk themselves; rather, it suggests that water utilities that are not required to implement these treatment practices may be at lower risk of drinking water lead exposure by nature of having less corrosive source waters or lower frequency of lead-bearing plumbing components. Thus, this analysis does not measure the effect of corrosion control treatment on blood lead levels—indeed, high risk

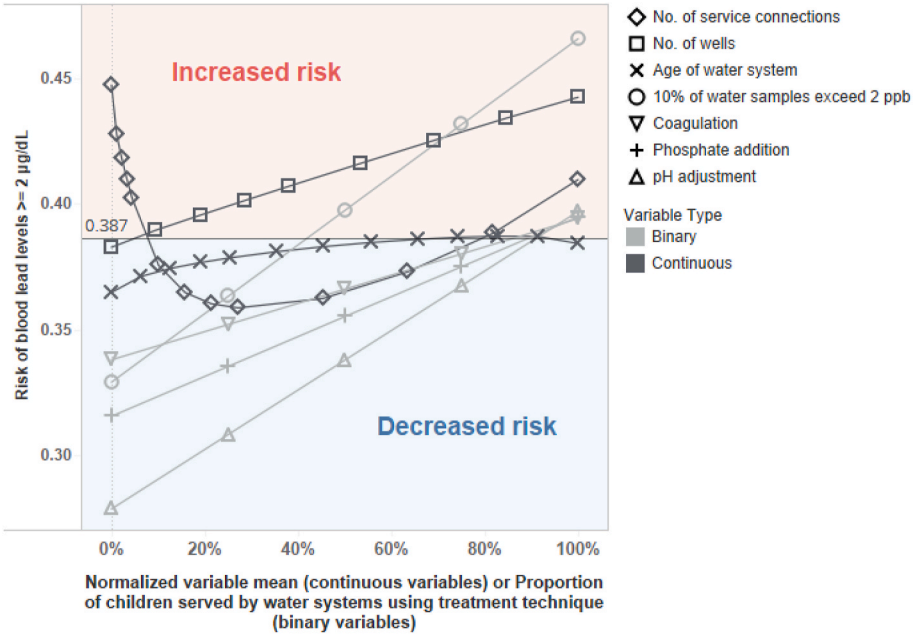

Fig. 6. Change in the risk (posterior probability) of blood lead levels ≥2 μg/dL among children served by community water systems by water system variable. The effect of each water system variable is shown while accounting for each of the other variables included in the network (Fig. 4). The x-axis shows either the normalized mean of the water system characteristic (for continuous variables) or the proportion of children served by systems using the specific treatment technique (for binary variables). Increased and decreased risks are shown relative to the prior probability in the county.

systems with very corrosive waters would certainly be at even higher risk without implementing corrosion control—but highlights the inability of corrosion control measures alone to completely eliminate lead exposures when corrosive source waters are used and suggests that the presence of such measures are an important attribute in assessing the overall lead exposure risk for a given system.

Increasing the number of service connections demonstrated a U-shaped effect, where very small (<286 service connections) and very large (>132,000 service connections) systems both were associated with greater risk of elevated blood lead levels. The reasons for this are not clearly elucidated by our model, but may be due to the unique challenges of each size category, such as water age concerns for large systems (Masters et al., 2015) and management and economic difficulties for small systems (Ford et al., 2005). Water system age had a small effect on blood lead levels, but children served by the newest systems in the county (<17 years old) were at slightly lower risk of elevated blood lead levels than those receiving water from older systems. Again, the reasons for this phenomenon are not captured by this analysis and are likely a complex interaction between scale-forming chemistry and system improvements (Cartier et al., 2013; Nguyen et al., 2011; Xie and Giammar, 2011).

Notably, children served by systems that exceeded the water lead threshold of 2 ppb in at least 10% of Lead and Copper Rule monitoring samples in the most recent year of sampling prior to the child's blood lead test also demonstrated an increase in the risk of blood lead levels ≥ 2 μg/dL. The probability of having a blood lead level ≥ 2 μg/dL increased to 46.6% among children served by systems that had exceeded this water lead threshold within two years of the blood lead test compared to 32.9% among children receiving water from systems that did not, representing an increase in risk of 42%. This exceedance variable also shared the most mutual information with the target out of all the water system characteristics included in the model (Fig. 5). Importantly, these systems would be considered fully compliant with all provisions of the Lead and Copper Rule.

Finally, a similar analysis of the effect of geographical clustering in our data set as described in Section 3.1.1 was performed for the 2 μg/dL target. When the 85% of children served by the two largest utilities were removed from the data set, the number of service connections, pH adjustment, number of wells, and the age of the water system continued to exhibit statistically significant mutual information with the target. The action level and trigger level remained insignificant, while children who were served by systems where 10% of the Lead and Copper Rule compliance monitoring samples exceeded 2 ppb continued to exhibit significantly greater risk of increased blood lead levels. From this we conclude that the selection of the water system variables in the model developed using the complete data set is representative of the nature of water lead exposures in the county as a whole, rather than only among the two largest systems.

3.2. Performance of Bayesian networks to predict blood lead levels

3.2.1. Model validation

On the full data set, the model achieved an area under the ROC curve of 80.47% which can be considered "good" overall predictive performance (Carter et al., 2016). During ten-fold cross-validation testing, the model structure achieved a comparable area under the ROC curve of 79.22%, indicating that the model is not subject to overfitting, with a tight confidence interval of ± 0.8% with different random partitions of the data into training and test sets (Figure S6). These scores indicate that the model could be expected to correctly rank the risk of a randomly chosen child with a blood lead level ≥2 μg/dL above a randomly chosen child with a blood lead level less than 2 μg/dL approximately 80% of the time on average (Hanley and McNeil, 1982).

The optimum decision threshold of the model, i.e., the probability used to determine if a child's blood lead level will be ≥2 μg/dL, can be

Table 1
Summary of model accuracy, sensitivity, and specificity for predicting whether a child's blood lead level will meet or exceed 2 μg/dL during cross validation with varying test thresholds.

Decision threshold	True positives	True negatives	Sensitivity	Specificity	Overall accuracy
0.1	15,320	6010	97%	24%	52%
0.2	14,256	11,908	91%	48%	64%
0.3	13,177	15,328	84%	61%	70%
0.4	11,774	17,864	75%	71%	73%
0.5	9945	20,166	63%	81%	74%
0.6	7575	22,235	48%	89%	73%
0.7	4787	23,785	30%	95%	70%
0.8	2256	24,639	14%	99%	66%
0.9	605	24,946	4%	100%	63%

selected to maximize the sensitivity, specificity, or overall accuracy of the model (Table 1). The decision threshold that maximizes the overall accuracy of the model is approximately 50%. That is, if a set of model inputs yielded a predicted risk of increased blood lead level of 50% or greater, then that child would be classified as having ≥ 2 μg/dL of blood lead, while if the calculated risk was less than 50%, the child would be considered to be below this level. At this threshold, the sensitivity of the model (true predicted positives/total actual positives) was 63% and the specificity (true predicted negatives/total actual negatives) was 81%. Thus, the model performed slightly better for predicting negative cases of subclinical blood lead levels than positive cases. The model's overall accuracy, which is the total number of correct predictions (true positives + true negatives) divided by the total number of cases, was 74%. However, in predicting childhood blood lead level risk, where the consequences of false positives are low compared to the consequences of false negatives, it may be desirable to sacrifice overall accuracy for improved sensitivity. As can be seen in Table 1, a lower test threshold of 30%, for example, would ensure that 84% of actual cases of blood lead levels ≥2 μg/dL are detected even though the specificity and accuracy at this threshold drop to 61% and 70%, respectively.

3.2.2. *Model implementation*

This model can be used to predict the risk of subclinical blood lead levels based on geographic, demographic, and water system characteristics to aid future blood lead exposure prevention programs. To illustrate this use, a new data set was compiled using each household in the original data currently served by a community water system. For each house, all the predictor variables shown in Fig. 4 were entered into the model. The child was assumed to be a boy aged 15–20 months (i.e., the demographic group with the highest blood lead levels on average). Once these inputs were specified, the model was run to predict the risk of having a blood lead level ≥2 μg/dL. In this way, the prediction provides an estimate of the highest risk areas for future testing.

The resulting calculated probabilities averaged across each Census block group and across each water system service area are shown in Fig. 7. The average predicted increase in blood lead level risk for children served by each water utility in our data set can be found in Table S8. Based on available information associated with each address, the model is capable of distinguishing spatial variations in the blood lead level risk associated with community water system attributes. Areas at higher risk of increased blood lead levels include parts of central Raleigh, where a cluster of block groups has a calculated risk of 30–50%, and several small unincorporated subdivisions with predicted probabilities exceeding 50%. Meanwhile, the cities of Apex, Cary, and Holly Springs exhibited a much lower predicted risk, generally less than 15%. The 23 water systems with the greatest overall average risk of blood lead levels ≥2 μg/dL (>30%) were all small groundwater systems with fewer than 500 connections. The majority of these systems implemented pH adjustment, but only one used phosphate corrosion inhibitors. Five water systems exceeded an average risk of blood lead levels ≥2 μg/dL of 50%. Overall, 64% of the children in our data set served by these 23 systems between 2002 and 2017 exhibited blood lead levels ≥2 μg/dL compared to a prevalence of only 25% among children served by water utilities in the lowest predicted risk category. Thus, Fig. 7 may help public health authorities, including the NC Childhood Lead Poisoning Prevention Program, prioritize areas for blood lead surveillance follow-up and alert drinking water utilities to potential water lead concerns.

4. Discussion

The Bayesian network models tested in this study identified a relationship between water system characteristics and blood lead levels at or above the regional median of 2 μg/dL. Elevated blood lead concentrations (≥5 μg/dL) were found to be more strongly predicted by socioeconomic and demographic factors, such as median household income and demographic composition of the Census block group of the child's address. These factors have previously been found to be correlated with the presence of lead paint and dust in homes, which are typically the primary cause of lead poisoning among children served by community water systems in the U.S. (Clark et al., 1985; Dixon et al., 2009; Lanphear et al., 2002). This finding does not necessarily indicate that no children served by community water systems experience blood lead levels above this level due to drinking water exposures. Indeed, the nature of lead release and prevalence of lead-bearing plumbing components ensure that isolated instances of blood lead levels exceeding 5 μg/dL attributable to community water systems may still occur even in well-managed systems.

The critical finding of this work was to identify the importance of drinking water characteristics toward predicting subclinical blood lead levels in children (i.e., blood lead levels at or above the regional median

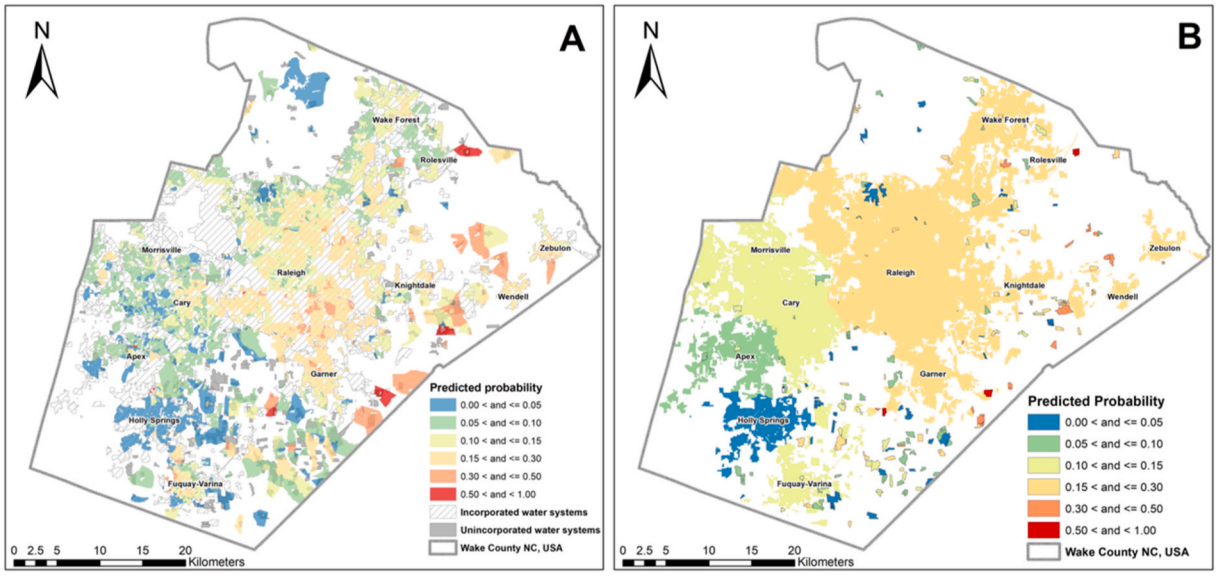

Fig. 7. Predicted probabilities of male children aged 15–19 months in Wake County having blood lead levels ≥2 μg/dL associated with exposures from community water systems. Panel A shows household level predictions averaged across Census blocks. Panel B shows household level predictions averaged across each approximate water system service area. Census blocks or water systems with fewer than three matched addresses were removed from the prediction.

of 2 μg/dL for Wake County, NC), even in community water systems not in violation of any regulatory provisions. Indeed, the machine-learned Bayesian networks indicated that multiple water system characteristics, including the system size, age and complexity, treatment characteristics, and 90th percentile water lead levels all significantly influence the probability of a child served by a community water system having a blood lead level ≥2 μg/dL. This supports previous research demonstrating that increased blood lead levels in children can be associated with water lead exposures even in populations served by community water systems considered not to have lead problems under the Lead and Copper Rule (Gleason et al., 2019; Katner et al., 2016; Lanphear et al., 2002). Although previous work on reducing water lead concentrations in community water systems in the U.S. has emphasized the lessons learned from high-profile cases such as the Washington D.C. and Flint, Michigan lead crises, which identified the catastrophic health effects associated with uncontrolled lead release in drinking water systems during major system-wide changes (Katner et al., 2016; Roy and Edwards, 2019), our findings suggest that the "lead crisis" in U.S. drinking water may at once be less overt and more prevalent.

This result has important implications for mitigating environmental lead exposures among children in the U.S. Lead prevention programs often focus on mitigation of paint and soil sources of lead, but potentially overlook the low, chronic contributions of water lead from community water systems when assessing the overall lead exposure risk profile. Such conditions can result in the "prevention paradox," where the largest burden of disease occurs in low to moderate risk categories while most prevention programs are focused on removing high exposure sources. Indeed, in 2012, the National Toxicology Program concluded that there was sufficient evidence that blood lead levels in children <5 μg/dL are associated with a broad range of adverse neurocognitive effects, including increased incidence of attention-related disorders, antisocial behaviors, decreased IQ, and poor performance in school (National Toxicology Program, 2012). Researchers have also identified a nonlinear effect between blood lead levels and IQ loss at low levels of exposure, indicating that the first slight increases of blood lead in infants and children have disproportionate impacts on neurological functioning (Canfield et al., 2003; Lanphear et al., 2005). As a result, the current policy focus on children with blood lead levels greater than 5 μg/dL alone is estimated to prevent only 20% of the IQ loss from lead among children in the U.S. (American Academy of Pediatrics Council on Environmental Health, 2016; Bellinger, 2012).

Similarly, the Lead and Copper Rule only requires the highest exposure risk sample sites to be selected based on locations that have lead service lines or lead solder rather than a representative sample to adequately assess population-wide lead exposure risk (Schwetschenau et al., 2020), which may still pose a significant health burden. Additionally, while current regulations are responsible for great reductions in water lead levels in the last three decades, the Bayesian network model for Wake County, NC suggested that the current Lead and Copper Rule action level and proposed trigger level are not sensitive enough to detect important variations in community water system lead exposure risk that are relevant to health today. Future studies are needed to identify a more appropriate threshold, but our findings suggest that a 90th percentile value of 2 ppb system-wide may be relevant to health on a population scale. Indeed, whether 10% of samples exceeded 2 ppb of lead in the most recent year of sampling prior to the blood test had a relatively large effect on predicting blood lead risk at or above the regional median, similar to the size of the effect of the year of house construction (Fig. 5), a commonly used predictor of lead exposure from household dust lead (Dixon et al., 2009; Gaitens et al., 2009; Gleason et al., 2019). In a multivariate regression of a large blood lead level data set in New Jersey, Gleason et al. (2019) showed that children served by water systems with a 90th percentile water lead concentration ≥2 ppb in the years 2000–2004 exhibited a 4% increase in geometric mean blood lead levels compared to children served by water systems with water lead levels <2 ppb. The importance of low water lead levels to blood lead risk also suggests that a lower water lead reporting limit should be enforced to ensure that samples in the 1–5 ppb range are accurately identified for all community water systems.

Even within the current Lead and Copper Rule framework, drinking water professionals and regulatory bodies may use the methods presented here and predictions of increased blood lead risk shown in Fig. 7 to proactively identify systems that may have higher lead exposure risk but may otherwise be in compliance. These systems also may need to identify plans to control lead in drinking water through lead service line replacement, corrosion control optimization, and distribution of water filters certified to remove lead. Further, these risks ought to be communicated to the public served by these systems so that individuals may also make informed decisions around mitigating their own and their children's lead exposure risk. Such efforts to continue to reduce water lead levels will help to offset lost economic productivity from lead exposure in the U.S. estimated to exceed $50 billion (Trasande and Liu, 2011).

Lastly, though these findings have important implications for water utilities and public health authorities in North Carolina among other regions in the U.S. with similar characteristics, this analysis provides a preliminary demonstration of the utility of machine-learned Bayesian networks for predicting lead exposure risks and may be improved through subsequent research and testing. First, the flattening of the data set for machine learning was a necessary manipulation of a complex relational database, but the model could potentially be improved using probabilistic relational modelling approaches (Getoor et al., 2007). Additionally, while machine learning algorithms are a powerful tool to identify the conditional probabilities embedded in a data set, how continuous variables are discretized depends on the decisions of the researcher and can have large impacts on the resulting model structure (Uusitalo, 2007). Thus, further study using the data set presented here using additional machine learning approaches along with traditional statistical approaches is ongoing. Finally, a heat map of increased blood lead levels associated with alternative exposures routes such as household paint and dust, private well water supplies (Macdonald Gibson et al., 2020), and leaded aviation gasoline (Miranda et al., 2011) would likely highlight additional areas of risk not predicted by our model in Fig. 7.

In summary, this work showed a significant relationship among water system characteristics and slight increases in blood lead levels in children in Wake County, NC. Public health authorities may use the machine-learning methods we present to help identify similar relationships among system-level drinking water characteristics and individual health outcomes. Although water lead exposures are particularly difficult to isolate and regulate, a lower health-based threshold and enhanced assessment of population-wide water lead exposure is critical to improve policies, prevention programs, and risk communication strategies that protect children from lead in drinking water.

5. Conclusion

This study is the first to link system-wide drinking water characteristics from multiple water utilities with individual health outcomes using machine learning techniques. We demonstrate that elevated blood lead levels (i.e., blood lead levels exceeding the CDC Reference Level of 5 μg/dL) are not generally associated with community drinking water systems in Wake County, NC, but that subclinical blood lead levels (i.e., blood lead levels ≥2 μg/dL) are strongly associated with community water system characteristics. Additionally, we demonstrate that machine learned Bayesian networks can accurately predict individual blood lead risk from system-wide water utility characteristics. Our results show that pairing public health and drinking water data using machine learning techniques can help to reveal the complex relationships between system-wide drinking water characteristics and public health outcomes.

Funding

This work was supported by the National Science Foundation [grant number 1832692] and the North Carolina Sea Grant/Water Resources Research Institute [grant number 19-04-W].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Mike Fisher, Ally Clonch, and Alex Shoaf at UNC Chapel Hill each made important contributions to the development of the data set. Thanks to Olga Naidenko at the Environmental Working Group for providing feedback on the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envres.2021.112146.

References

Adams, H., 1859. On the action of water on lead pipes, and the diseases proceeding from it. In: Kirkwood, J.P. (Ed.), Collection of Reports, (Condensed), and Opinions of Chemists in Regard to the Use of Lead Pipe for Service Pipe. Hosford and Co., New York, NY, pp. 126–179 the Distribution of Water for the Supply of Cities.

American Academy of Pediatrics Council on Environmental Health, 2016. Prevention of childhood lead toxicity. Pediatrics 138, 1–15. https://doi.org/10.1542/peds.2016-1493.

Angelon-Gaetz, K., Newman Chelminski, A., 2018. Trends in lead poisoning prevention data for children aged <6 Years in North Carolina. N. C. Med. J. 79, 339–342.

ATSDR, 2007. Toxicological Profile for Lead. Agency for Toxic Substances and Disease Registry (ATSDR). U.S. Department of Health and Human Services, Atlanta, GA.

Barker, G.C., 2013. A Risk Assessment Model for Enterotoxigenic Staphylococcus aureus in Pasteurized Milk : A Potential Route to Source-Level Inference 33. https://doi.org/10.1111/j.1539-6924.2011.01667.x.

Bayesia, S.A.S., 2021. R2-GenOpt [WWW Document]. BayesiaLab User Guid. https://library.bayesia.com/articles/#!bayesialab-knowledge-hub/discretization-r2-genopt/q/r2-genopt/qid/3627/qp/1. accessed 9.19.21.

Bellinger, D., Leviton, A., Waternaux, C., Needleman, H., Rabinowitz, M., 1987. Longitudinal analyses of prenatal and postnatal lead exposure and early cognitive development. 3The New Engl. J. Med. 316, 1037–1043.

Bellinger, D.C., 2012. A strategy for comparing the contributions of environmental chemicals and other risk factors to neurodevelopment of children. Environ. Health Perspect. 120, 501–507. https://doi.org/10.1289/ehp.1104170.

Canfield, R.L., Henderson, C.R.J., Cory-Slechta, D.A., Cox, C., Jusko, T.A., Lanphear, B.P., 2003. Intellectual impairment in children with blood lead concentrations below 10 μg per deciliter. N. Engl. J. Med. 348, 1517–1526.

Carter, J.V., Pan, J., Rai, S.N., Galandiuk, S., 2016. ROC-ing along : evaluation and interpretation of receiver operating characteristic curves. Surgery 159, 1638–1645. https://doi.org/10.1016/j.surg.2015.12.029.

Cartier, C., Doré, E., Laroche, L., Nour, S., Edwards, M., Prévost, M., 2013. Impact of treatment on Pb release from full and partially replaced harvested Lead Service Lines (LSLs). Water Res. 47, 661–671. https://doi.org/10.1016/j.watres.2012.10.033.

Clark, C.S., Bornschein, R.L., Succop, P., Hee, S.S.Q., Hammond, P.B., Peace, B., 1985. Condition and type of housing as an indicator of potential environmental lead exposure and pediatric blood lead levels. Environ. Res. 38, 46–53. https://doi.org/10.1016/0013-9351(85)90071-4.

Conrady, S., Jouffe, L., 2015. Bayesian Networks and Bayesia Lab: A Practical Inroduction for Researchers (Bayesia USA, Franklin, TN).

Cornwell, D.A., Brown, R.A., Via, S.H., 2016. National survey of lead service line occurrence. J. Am. Water Works Assoc. 108, E182–E191. https://doi.org/10.5942/jawwa.2016.108.0086.

Del Toral, M.A., Porter, A., Schock, M.R., 2013. Detection and evaluation of elevated lead release from service lines: a field study. Environ. Sci. Technol. 47, 9300–9307. https://doi.org/10.1021/es4003636.

Dignam, T., Kaufmann, R.B., Lestourgeon, L., Brown, M.J., 2019. Control of lead sources in the United States, 1970-2017: public health progress and current challenges to eliminating lead exposure. J. Publ. Health Manag. Pract. 25, S13–S22. https://doi.org/10.1097/PHH.0000000000000889.

Dignam, T.A., Lojo, J., Meyer, P.A., Norman, E., Sayre, A., Flanders, W.D., 2008. Reduction of elevated blood lead levels in children in North Carolina and Vermont, 1996-1999. Environ. Health Perspect. 116, 981–985. https://doi.org/10.1289/ehp.10548.

Dixon, S.L., Gaitens, J.M., Jacobs, D.E., Strauss, W., Nagaraja, J., Pivetz, T., Wilson, J.W., Ashley, P.J., 2009. Exposure of U.S. children to residential dust lead, 1999-2004: II. The contribution of lead-contaminated dust to children's blood lead levels. Environ. Health Perspect. 117, 468–474. https://doi.org/10.1289/ehp.11918.

EWG, 2019. Tap Water Database [WWW Document]. https://www.ewg.org/tapwater/. accessed 7.6.20.

Fasaee, M.A.K., Berglund, E., Pieper, K.J., Ling, E., Benham, B., Edwards, M., 2021. Developing a framework for classifying water lead levels at private drinking water systems: a Bayesian Belief Network approach. Water Res. 189, 116641. https://doi.org/10.1016/j.watres.2020.116641.

Ford, T., Rupp, G., Butterfield, P., Camper, A., 2005. Protecting Public Health in Small Water Systems: Report of an International Colloquium. Montana Water Center, Montana State University, Bozeman, Montana.

Francis, R.A., Guikema, S.D., Henneman, L., 2014. Bayesian Belief Networks for predicting drinking water distribution system pipe breaks. Reliab. Eng. Syst. Saf. 130, 1–11. https://doi.org/10.1016/j.ress.2014.04.024.

Friedman, N., 1997. Learning belief networks in the presence of missing values and hidden variables. In: Fisher, D.H. (Ed.), ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc.

Gaitens, J.M., Dixon, S.L., Jacobs, D.E., Nagaraja, J., Strauss, W., Wilson, J.W., Ashley, P.J., 2009. Exposure of U.S. children to residential dust lead, 1999-2004: I. Housing and demographic factors. Environ. Health Perspect. 117, 461–467. https://doi.org/10.1289/ehp.11917.

Getoor, L., Friedman, N., Koller, D., Pfeffer, A., Taskar, B., 2007. Probabilistic relational models. In: Getoor, L., Taskar, B. (Eds.), Introduction to Statistical Relational Learning. MIT Press, Cambridge, MA. https://doi.org/10.7551/mitpress/7432.001.0001.

Gleason, J.A., Nanavaty, J.V., Fagliano, J.A., 2019. Drinking water lead and socioeconomic factors as predictors of blood lead levels in New Jersey ' s children between two time periods. Environ. Res. 169, 409–416. https://doi.org/10.1016/j.envres.2018.11.016.

Goldszmidt, M., 2011. Bayesian network classifiers. Wiley Encycl. Oper. Res. Manag. Sci. 163, 131–163. https://doi.org/10.1002/9780470400531.eorms0099.

Gray, W.B., 1916. Lead data. In: Gray's Plumbing Design and Installation; a Veritable Encyclopedia of Modern Practice Based on Work Done by the Author and Other Experts in Every Branch of the Plumbing and Allied Trades and Covering Approved Practice in Every Part of the Country. David Williams, New York, NY, pp. 105–113.

Guzella, T.S., Caminhas, W.M., 2009. A review of machine learning approaches to Spam filtering. Expert Syst. Appl. 36, 10206–10222. https://doi.org/10.1016/j.eswa.2009.02.037.

Haley, V.B., Talbot, T.O., 2004. Seasonality and trend in blood lead levels of New York State children. BMC Pediatr. 4, 1–5. https://doi.org/10.1186/1471-2431-4-8.

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36. https://doi.org/10.1148/radiology.143.1.7063747.

Hodge, T.A., 1981. Vitruvius, lead pipes and lead poisoning. Am. J. Archaeol. 85, 486–491.

Hu, Q., Whitney, H.M., Giger, M.L., 2020. A deep learning methodology for improved breast cancer diagnosis using multiparametric MRI. Sci. Rep. 10, 1–11. https://doi.org/10.1038/s41598-020-67441-4.

Katner, A., Pieper, K.J., Lambrinidou, Y., Brown, K., Hu, C.-Y.Y., Mielke, H.W., Edwards, M.A., 2016. Weaknesses in federal drinking water regulations and public health policies that impede lead poisoning prevention and environmental justice. Environ. Justice 9, 109–117. https://doi.org/10.1089/env.2016.0012.

Kotsiantis, S.B., 2007. Supervised machine learning: a review of classification techniques. In: Maglogiannis (Ed.), Emerging Artificial Intelligence Applications in Computer Engineering. IOS Press, pp. 3–24.

Kramer, S., Lavrač, N., Flach, P., 2001. Propositionalization approaches to relational data mining. In: Džeroski, S. (Ed.), Relational Data Mining. Springer-Verlag, Berlin, pp. 262–291. https://doi.org/10.1007/978-3-662-04599-2_11.

Lambrinidou, Y., Triantafyllidou, S., Edwards, M., 2010. Failing our children: lead in U.S. school drinking water. New Solut. 20, 25–47. https://doi.org/10.2190/NS.022010eov.

Lanphear, B.P., Hornung, R., Ho, M., Howard, C.R., Eberle, S., Knauf, K., 2002. Environmental lead exposure during early childhood. J. Pediatr. 140, 40–47. https://doi.org/10.1067/mpd.2002.120513.

Lanphear, B.P., Hornung, R., Khoury, J., Yolton, K., Baghurst, P., Bellinger, D.C., Canfield, R.L., Dietrich, K.N., Bornschein, R., Greene, T., Rothenberg, S.J., Needleman, H.L., Schnaas, L., Wasserman, G., Graziano, J., Roberts, R., 2005. Low-level environmental lead exposure and children's intellectual function: an international pooled analysis. Environ. Health Perspect. 113, 894–899. https://doi.org/10.1289/ehp.7688.

Lanphear, B.P., Rauch, S., Auinger, P., Allen, R.W., Hornung, R.W., 2018. Low-level lead exposure and mortality in US adults: a population-based cohort study. Lancet Public Heal 3, e177–e184. https://doi.org/10.1016/S2468-2667(18)30025-2.

Lee, J., Henning, R., Cherniack, M., 2019. Correction workers' burnout and outcomes: a bayesian network approach. Int. J. Environ. Res. Publ. Health 16. https://doi.org/10.3390/ijerph16020282.

Lessler, M.A., 1988. Lead and lead poisoning from antiquity to modern times lead. Ohio J. Sci. 88, 78–84.

Li, W.T., Ma, J., Shende, N., Castaneda, G., Chakladar, J., Tsai, J.C., Apostol, L., Honda, C.O., Xu, J., Wong, L.M., Zhang, T., Lee, A., Gnanasekar, A., Honda, T.K., Kuo, S.Z., Yu, M.A., Chang, E.Y., Rajasekaran, M.R., Ongkeko, W.M., 2020. Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. BMC Med. Inf. Decis. Making 20, 1–13. https://doi.org/10.1186/s12911-020-01266-z.

Lobo, G.P., Laraway, J., Gadgil, A.J., 2021. Identifying schools at high-risk for elevated lead in drinking water using only publicly available data. Sci. Total Environ. 803, 150046. https://doi.org/10.1016/j.scitotenv.2021.150046.
Maas, R.P., Patch, S.C., Morgan, D.M., Pandolfo, T.J., 2005. Reducing lead exposure from drinking water: recent history and current status. Publ. Health Rep. 120, 316–321. https://doi.org/10.1177/003335490512000317.
Macdonald Gibson, J., Fisher, M., Clonch, A., Macdonald, J.M., Cook, P.J., 2020. Children drinking private well water have higher blood lead than those with city water. Proc. Natl. Acad. Sci. Unit. States Am. https://doi.org/10.1073/pnas.2002729117.
Maier, M., Marazopoulou, K., Arbour, D., Jensen, D., 2013. Flattening network data for causal discovery: what could go wrong?. In: 5th Work. Inf. Networks.
Masters, S., Parks, J., Atassi, A., Edwards, M.A., 2015. Distribution system water age can create premise plumbing corrosion hotspots. Environ. Monit. Assess. 187 https://doi.org/10.1007/s10661-015-4747-4.
McDonald, J.H., 2014. G-test of goodness-of-fit. In: Handbook of Biological Statistics. Sparky House Publishing, Baltimore, pp. 53–58.
McMichael, A.J., Baghurst, P.A., Wigg, N.R., Vimpani, G.V., Robertson, E.F., Roberts, R. J., 1988. Port pirie cohort study: environmental exposure to lead and children's abilities at the age of four years. N. Engl. J. Med. 319, 468–475.
Miranda, M.L., Anthopolos, R., Hastings, D., 2011. A geospatial analysis of the effects of aviation gasoline on childhood blood lead levels. Environ. Health Perspect. 119, 1513–1516. https://doi.org/10.1289/ehp.1003231.
National Toxicology Program, 2012. NTP Monograph on Health Effects of Low-Level Lead. National Institute of Environmental Health Sciences, Research Triangle Park, NC.
NCDEQ, 2020. Drinking Water Watch [WWW Document]. https://www.pwss.enr.state.nc.us/NCDWW2/. accessed 7.6.20.
NCDHHS, 2021. Childhood Lead Poisoning Prevention Program. Data [WWW Document]. https://ehs.ncpublichealth.com/hhccehb/cehu/lead/data.htm. accessed 9.23.21.
Needleman, H.L., Schell, A., Bellinger, D., Leviton, A., Allred, E.N., 1990. The long-term effects of exposure to low doses of lead in childhood: an 11-year follow-up report. N. Engl. J. Med. 322, 83–88.
Nguyen, C.K., Clark, B.N., Stone, K.R., Edwards, M.A., 2011. Role of chloride, sulfate, and alkalinity on galvanic lead corrosion. Corrosion 67. https://doi.org/10.5006/1.3600449, 065005-1-065005–9.
NRDC, 2021. Lead Pipes Are Widespread and Used in Every State. Natural Resources Defense Council, New York, NY.
Olsen, C.R., Mentz, R.J., Anstrom, K.J., Page, D., Patel, P.A., 2020. Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure: machine learning in heart failure. Am. Heart J. 229, 1–17. https://doi.org/10.1016/j.ahj.2020.07.009.
Orme-zavaleta, J., Jorgensen, J., Ambrosio, B.D., Altendorf, E., Rossignol, P.A., 2006. Discovering Spatio-Temporal Models of the Spread of West Nile Virus, vol. 26. https://doi.org/10.1111/j.1539-6924.2006.00738.x.
Poropat, A.E., Laidlaw, M.A.S., Lanphear, B., Ball, A., Mielke, H.W., 2018. Blood lead and preeclampsia: a meta-analysis and review of implications. Environ. Res. 160, 12–19. https://doi.org/10.1016/j.envres.2017.09.014.
Potash, E., Brew, J., Loewi, A., Majumdar, S., Reece, A., Walsh, J., Rozier, E., Jorgensen, E., Mansour, R., Ghani, R., 2015. Predictive modeling for public health: preventing childhood lead poisoning. In: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 2015-Augus, pp. 2039–2047. https://doi.org/10.1145/2783258.2788629.
Rabin, R., 2008. The lead industry and lead water pipes "A modest campaign. Am. J. Publ. Health 98, 1584–1592. https://doi.org/10.2105/AJPH.2007.113555.
Redmon, J.H., Gibson, J.M.D., Woodward, K.P., Aceituno, A.M., Levine, K.E., 2018. Safeguarding children's health: time to enact a health-based standard and comprehensive testing, mitigation, and communication protocol for lead in drinking water. N. C. Med. J. 79, 313–317. https://doi.org/10.18043/ncm.79.5.313.
Riblet, C., Deshommes, E., Laroche, L., Prévost, M., 2019. True exposure to lead at the tap: insights from proportional sampling, regulated sampling and water use monitoring. Water Res. 156, 327–336. https://doi.org/10.1016/j.watres.2019.03.005.
Roy, S., Edwards, M.A., 2019. Preventing another lead (Pb) in drinking water crisis: lessons from the Washington D.C. and Flint MI contamination events. Curr. Opin. Environ. Sci. Heal. 7, 34–44. https://doi.org/10.1016/j.coesh.2018.10.002.
Schwetschenau, S., Small, M.J., Vanbriesen, J.M., 2020. Using compliance data to understand uncertainty in drinking water lead levels in southwestern Pennsylvania. Environ. Sci. Technol. 54, 8857–8867. https://doi.org/10.1021/acs.est.9b07303.
Sebastiani, P., Perls, T.T., 2008. Complex genetic models. In: Pourret, O., Naim, P., Marcot, B. (Eds.), Bayesian Networks: A Practical Guide to Applications. John Wiley & Sons, Ltd, Chichester, England, pp. 53–72. https://doi.org/10.1002/9780470994559.ch4.
Stark, A.D., Quah, F., Meigs, J.W., Delouise, R., Stark, A.D., 1982. Relationship of Sociodemographic Factors to Blood Lead Concentrations in New Haven Children, pp. 133–139.
Tang, Z., Hong, S., Xiao, W., Taylor, J., 2006. Impacts of blending ground, surface, and saline waters on lead release in drinking water distribution systems. Water Res. 40, 943–950. https://doi.org/10.1016/j.watres.2005.12.028.
Trasande, L., Liu, Y., 2011. Reducing the staggering costs of environmental disease in children, estimated at $76.6 billion in 2008. Health Aff. 30, 863–870. https://doi.org/10.1377/hlthaff.2010.1239.
Triantafyllidou, S., Edwards, M., 2012. Lead (Pb) in tap water and in blood: implications for lead exposure in the United States. Crit. Rev. Environ. Sci. Technol. 42, 1297–1352. https://doi.org/10.1080/10643389.2011.556556.
Trueman, B.F., Camara, E., Gagnon, G.A., 2016. Evaluating the effects of full and partial lead service line replacement on lead levels in drinking water. Environ. Sci. Technol. 50 https://doi.org/10.1021/acs.est.6b01912.
USEPA, 1991. Lead and copper Rule. Fed. Regist. 56, 26460–26464.
USEPA, 2000. National primary drinking water regulations for lead and copper. Fed. Regist. 65, 1950–2015.
USEPA, 2004. National primary drinking water regulations: minor corrections and clarification to drinking water regulations; national primary drinking water regulations for lead and copper. Fed. Regist 69, 38850–38857.
USEPA, 2007. National primary drinking water regulations for lead and copper: short-term regulatory revisions and clari cations. Fed. Regist. 72, 57781–57820.
USEPA, 2019. National Primary Drinking Water Regulations: Proposed Lead and Copper Rule Revisions. https://doi.org/10.1016/0196-335x(80)90058-8. Federal Register.
Uusitalo, L., 2007. Advantages and challenges of Bayesian networks in environmental modelling. Ecol. Model. 203, 312–318. https://doi.org/10.1016/j.ecolmodel.2006.11.033.
Vuorinen, H.S., Juuti, P.S., Katko, T.S., 2019. Safety of Lead Water Pipes: History and Present, Water Services Management and Governance. Lessons for a Sustainable Future. https://doi.org/10.2166/9781780400730.
Whitehead, L.S., Buchanan, S.D., 2019. Childhood lead poisoning: a perpetual environmental justice issue? J. Publ. Health Manag. Pract. 25, S115–S120. https://doi.org/10.1097/PHH.0000000000000891.
Xie, Y., Giammar, D.E., 2011. Effects of flow and water chemistry on lead release rates from pipe scales. Water Res. 45, 6525–6534. https://doi.org/10.1016/j.watres.2011.09.050.
Yiin, L., Rhoads, G.G., Lioyl, P.J., 2000. Seasonal influences on childhood lead exposure. Environ. Health Perspect. 108, 177–182.
Zabinski, J.W., Garcia-Vargas, G., Rubio-Andrade, M., Fry, R.C., MacDonald Gibson, J., 2016. Advancing dose-response assessment methods for environmental regulatory impact analysis: a bayesian belief network approach applied to inorganic arsenic. Environ. Sci. Technol. Lett. 3, 200–204. https://doi.org/10.1021/acs.estlett.6b00076.Advancing.
Zabinski, J.W., Pieper, K.J., Gibson, J.M., 2018. A bayesian belief network model assessing the risk to wastewater workers of contracting Ebola virus disease during an outbreak. Risk Anal. 38, 376–391. https://doi.org/10.1111/risa.12827.