Published in final edited form as:

J Health Care Poor Underserved. 2013 February ; 24(1 0): 153–171. doi:10.1353/hpu.2013.0046.

Analysis of an Environmental Exposure Health Questionnaire in a Metropolitan Minority Population Utilizing Logistic Regression and Support Vector Machines

Chau-Kuang Chen, EdD, Michelle Bruce, MD, MSPH, Lauren Tyler, BS, Claudine Brown, MSPH, Angelica Garrett, MD, Susan Goggins, MD, Brandy Lewis-Polite, MD, Mirabel L Weriwoh, MD, MSPH, Paul D. Juarez, PhD, Darryl B. Hood, PhD, and Tyler Skelton, MS
Department of Institutional Research at Meharry Medical College (MMC) [C-KC]; Department of Internal Medicine , MMC [MB]; the Department of Biology, Tennessee State University [LT]; the School of Medicine at MMC [AG, SG, B L-P]; the School of Graduate Studies and Research, MMC [CB, MLW]; the Department of Family and Community Medicine, MMC [PDJ, TS]; and the Department of Neuroscience and Pharmacology, NIMHD Health Disparities Research Center of Excellence, Environmental Health Disparities and Medicine, Center for Molecular and Behavioral Neuroscience, MMC [DBH].

Abstract

The goal of this study was to analyze a 54-item instrument for assessment of perception of exposure to environmental contaminants within the context of the built environment, or *exposome*. This *exposome* was defined in five domains to include 1) home and hobby, 2) school, 3) community, 4) occupation, and 5) exposure history. Interviews were conducted with child-bearing-age minority women at Metro Nashville General Hospital at Meharry Medical College. Data were analyzed utilizing DTReg software for Support Vector Machine (SVM) modeling followed by an SPSS package for a logistic regression model. The target (outcome) variable of interest was respondent's residence by ZIP code. The results demonstrate that the rank order of important variables with respect to SVM modeling *versus* traditional logistic regression models is almost identical. This is the first study documenting that SVM analysis has discriminate power for determination of higher-ordered spatial relationships on an environmental exposure history questionnaire.

Keywords

Support vector machines; questionnaire validation; environmental exposure; minority populations; public participatory-geographical information systems

The socioeconomic circumstances of people and the places where they live work and play strongly influence their health.¹ Healthy homes are essential to a healthy community and

serve to protect families from exposure to potential environmental contaminants (such as chemicals and allergens) thereby preventing exposure-related phenomena. In contrast, inadequate housing contributes to infectious and chronic disease and injuries and has been shown to affect child development adversely.¹ The definition of inadequate housing is related to the basic structure and systems of a housing unit, whereas the definition of unhealthy housing is related to exposure to environmental pollutants and other toxicants.¹ The Centers for Disease Control and Prevention (CDC) has defined unhealthy housing as housing with characteristics that might negatively affect the health of its occupants, including rodents, water leaks, peeling paint in homes built before 1978, and absence of a working smoke detector.

The most recent CDC report concludes that efforts to decrease disparities in access to healthy housing will have the immediate effect of decreasing disparities in health status.¹ Among the approximately 110 million housing units in the United States, approximately 5.8 million are classified as inadequate, and 23.4 million are considered unhealthy. Inadequate and unhealthy housing disproportionately affect populations that have the fewest resources (e.g., people with low incomes and limited education). Studies reveal that certain ethnic and racial populations (such as African Americans), have a greater likelihood of residing in areas in which there is poor housing and poor air quality primarily defined in terms of fine particles (PM_{2.5}).² Minority populations are especially vulnerable and susceptible to exposure to PM_{2.5} as a result of their homes being located in close proximity to the emissions from smokestacks and vehicles (heavy traffic volumes).^{3,4} Historical studies have shown that vulnerable segments of minority populations, such as pregnant women and young children, are particularly susceptible to the effects of PM_{2.5} air pollution.^{5–8} These facts illustrate the need to develop survey instruments for the physician to have at his fingertips when taking initial environmental exposure histories. This is particularly relevant for physicians practicing in indigent care hospitals where many health problems plague urban minority populations with low socioeconomic status.

Recent reports on exposure during pregnancy to components of air pollution have revealed deleterious effects on development of the fetus. High levels of components of photochemical smog, including carbon monoxide (CO), nitrogen dioxide (NO₂), and ambient particulate matter (PM_{2.5}) have been associated with very low and low birth weight, preterm birth, and infant mortality.^{9,10} Particulate matter is material that is suspended in the air in the form of minute solid particles or liquid droplets as an atmospheric pollutant. Such environmental contaminants have also been associated with significant differences in biparietal diameter and head circumference both measured during pregnancy and at birth.^{11,12} *In utero* exposure to polycyclic aromatic hydrocarbons (PAHs) during pregnancy has been shown to correlate with impaired cortical function and cognitive developmental delay.^{13–17}

A recent epidemiologic study reported that use of gas appliances and increased NO₂ in the home during the first three months of life are associated with decreased cognitive test scores and increased inattention at four years of age.¹⁸ In a separate study, Suglia and colleagues estimated lifetime residential exposure to carbon black, a proxy for traffic-related PM, among 8–11 year old children with an associated decline in performance on intelligence and memory tasks as a function of increasing carbon black levels.¹⁹ Interestingly, autism

spectrum disorder has been associated with estimated regional concentrations of hazardous air pollutants, including arsenic and nickel, and with diesel PM exposure in early childhood.²⁰ Thus, there is an emerging literature suggesting that nearby roadway traffic-related air pollutants, possibly influenced by specific components such as particulate matter or PAHs, adversely affect neurodevelopmental processes.

In the present study conducted out of Metro Nashville General Hospital at Meharry Medical College, we have used support vector machines (SVM) to analyze responses to a 54-item environmental exposure questionnaire towards validation of the instrument for future use in recruiting pregnant women for prospective longitudinal studies. Support vector machines are a relatively new classification or prediction method developed by Cortes and Vapnik²¹ in the 1990s as a result of the collaboration between the statistical and the machine-learning research communities. Support vector machines attempt to classify datasets by finding a separating boundary, referred to as a *hyperplane*. The main advantage of the SVM is that it can, with relative ease, overcome the high dimensionality problem, i.e., the problem that arises when there is a large number of input variables relative to the number of available observations.²² Further, because the SVM approach is data-driven and possible without a theoretical framework, it is believed to have important discriminative power for classification, especially with datasets where sample sizes are small.²³ The technique has recently been used to improve methodology for detecting diseases in clinical settings.^{24,25} The SVM analysis of responses from this study will serve as a point of reference for recruitment of a prospective cohort of pregnant African American women living in an inner city environment to quantify the frequency of exposure related adverse pregnancy outcomes. (841 words)

Methods

Demonstration of similarity of demographics between inner-city atlanta area (grady Memorial Hospital) and inner-city nashville area (Metro nashville general Hospital at Meharry Medical College) as a means to validate our 54-item instrument

All procedures in accordance with administration of the 54-item instrument within the subspecialty clinics at Nashville General Hospital at Meharry Medical College were submitted to the Meharry Medical College Institutional Review Board (IRB). The Institutional Review Board determined that the project was exempt, based on category 45 CFR 46.101.b (2) of the federal regulations concerning the use of survey procedures when the information is recorded in such a manner that subjects cannot be identified. No consent form was needed and the IRB application was approved (IRB number 051115DBH195-11).

Our 54-item environmental exposure instrument is an adaptation of the Agency for Toxic Substances and Disease Registry (ATSDR) environmental exposure history questionnaire (http://www.atsdr.cdc.gov/hec/csem/exphistory/docs/exposure_history.pdf.) and therefore required validation and calibration. As a means of validating our instrument, we sought to determine the similarity of demographic characteristics between inner-city Atlanta, Georgia and inner-city Nashville, Tennessee based on 2000 and 2010 census data. For additional clarity, we also determined the similarity of demographic characteristics between the Grady Memorial Hospital area and the Metro Nashville General Hospital at Meharry area. These

data are shown in Table 1 and indicate that the respondent population at Grady Memorial Hospital in inner-city Atlanta and the respondent population at Metro Nashville General Hospital at Meharry are very similar in terms of the Black and/or African American representation. The percent differences in 2012 demographic data for Atlanta compared with Nashville in favor of Atlanta (16.4%) is offset by the robust African American demographic (97.3%) in ZIP code 37208 (where Metro Nashville General Hospital at Meharry is located). The Grady Memorial Hospital area's African American population represents 76.3% of the total population in the area. There are similar percent differences between the Atlanta Grady Memorial Hospital area and the Metro Nashville General Hospital at Meharry area in educational attainment, sex, and school enrollment. All of these indices proved important as a part of the calibration-validation process regarding our 54-item instrument.

Study subjects

A total of 187 child-bearing-age (18–35) African American or Hispanic women were interviewed when presenting for services at the Obstetrics and Gynecology, Family and Community Medicine, or Pediatrics subspecialty clinics at Metropolitan Nashville General Hospital at Meharry Medical College.

Personal interview

The 54-item environmental exposure questionnaire was administered to consenting eligible child-bearing-age (18–35 years) African American and Hispanic women. The survey was partitioned into five domains including 1) home and hobby, 2) school, 3) community, 4) occupation, and 5) exposure history. Demographic information pertaining to the woman's age and education level were recorded, as was the ZIP code in which she primarily resided. This 54-item questionnaire was modeled from both the CDC and the ATSDR environmental exposure history questionnaires.²⁶

Exposed vs. control dichotomization of ZIP codes

Based on the proximity of EPA reporting sites with emissions to Meharry Medical College-respondents that resided in ZIP codes 37206, 37207, 37208 and 37209 were assigned to the exposed group. Respondents residing in any other ZIP codes (37210, 36606, 37011, 37013, 37062, 37075, 37076, 37082, 37086, 37087, 37115, 37219, 37138, 37203, 37210, 37211, 37212, 37213, 37214, 37215, 37216, 37217, 37218, 37219, 37220, 37221, 37228, 37235, and 38556) were assigned to the control group.

Overview of the SVM classifier

The SVM algorithm is a supervised machine learning method which has demonstrated its ability to solve complex classification problems, especially in the field of bioinformatics.^{27,28} The SVM model is operated on the principle of structural risk minimization.²⁹ It is designed to minimize true risk of misclassifying examples during the model training. It has its advantage in the practical application for small sample and generalization because of structural risk minimization.^{29–31}

Unlike many other more popular data analysis tools, the SVM algorithm is data-driven and model-free and, as such, may have much discriminative power for classification in instances

where many explanatory variables are used and the sample size is relatively small. Support vector machines establishes sophisticated models by constructing a multidimensional hyperplane optimally discriminating (maximizing the margin separation) between two different classes.³² The algorithm achieves its high discriminative power by using special non-linear functions called kernel functions to transform the original data into a high-dimensional space.³³ Justified by Cover's theorem, any dataset can be separated out if the data dimension grows.

The basic role of the kernel functions is to calculate inner product values through a transformation in high-dimensional feature space, and ultimately maximize the margin of separation to yield high accuracy of data separation. All four kernel functions are involved in the model construction: linear, radial basis function (RBF), polynomial, and sigmoid. Each SVM analysis is equipped with only one kernel function to fit a hyperplane between different classes of examples. The kernels are mathematically defined as: (1) Linear kernel, $k(u, v) = u \cdot v$; (2) RBF kernel, $k(u, v) = \exp(-\gamma \|u - v\|^2)$; (3) Sigmoid kernel, $k(u, v) = \tanh(\gamma u \cdot v + \text{constant})$; and (4) Polynomial kernel, $k(u, v) = (\gamma u \cdot v + 1)^d$, where u and v are input vectors: $\exp$ is the base of the natural logarithm, and d represents the number of degrees in polynomial kernel.²¹ The most suitable kernel function can be chosen based on the sensitivity, specificity, and overall accuracy of the function to classify accurately the outcome categories.

In order to yield the best performance model, SVM adopts the iterative process of grid-search (searches values of each parameter across the specified search range) and pattern search (searches the center and makes trial steps in each direction for each parameter) to find the optimal value of parameters.³⁴ All pairs of optimal parameters such as kernel function weight (gamma “ γ ”) and the cost of error (“ C ”) in the system are being tried, and the best pair is generated.²¹ The SVM model also implements m -fold cross-verification to obtain the reliable estimation of model error, where for each of m experiments, the system uses $m-1$ folds for training and the remaining one for testing. In the validation process, the average of model errors for distinct categories is calculated to evaluate the model accuracy. The advantage of m -fold cross-verification is that all examples in the dataset are eventually used for both training and testing to avoid sampling bias problems.

Data analysis

A database was created from the collected responses. Entries with missing data points were excluded from the analysis and key domains of interest were identified. The overall response rate to the 54-item instrument was (94.3%). Our method converted the variable which spoke to the respondent's neighborhood from a categorical variable (Rural, Suburban and Urban) into a dichotomous variable by combining the urban and suburban fields for the sake of parity. Respondents' ZIP codes were dichotomized to form an exposure and control group with persons residing in the 37206, 37207, 37208 and 37209 ZIP codes being assigned to the exposed group based on the number of EPA reporting industrial facilities and sites coupled with their proximity to Meharry Medical College. Respondents residing in ZIP codes (37210, 36606, 37011, 03713, 37062, 37075, 37076, 37082, 37086, 37087, 37115, 37219, 37138, 37203, 37210, 37211, 37212, 37213, 37214, 37215, 37216, 37217, 37218, 37219,

37220, 37221, 37228, 37235, and 38556) outside of this area, were assigned to the control group. Data analysis was performed on the dataset using DTReg software for Support Vector Machine (Sherrod, PH., 2008, Digital Tree Regression (DTREG) (Version 4.0) [software]. Brentwood, Tennessee) and SPSS version 17 for the logistic regression model (Statistical Product and Service Solutions (SPSS) (Version 17) [software] Armonk, New York).

Six steps were systematically adopted to yield the most suitable SVM and logistic regression models to accurately classify respondents living in distinct ZIP code groups.

Step 1 involved generating a research dataset consisting of one dependent variable and multiple independent variables collected from respondent's questionnaires.

In Step 2, the underlying models were developed by the difficult selection of kernel functions for the SVM. Thus, our only strategy was performed a trial and error process by applying all kernel functions. For SVM classifier, the following four kernel functions were readily available for model construction: linear, radial basis function (RBF), polynomial, and sigmoid.

In Steps 3 and 4, all possible SVM candidate models were trained and tested to achieve minimal classification error by means of adjusting parameters and performing cross-validation. For the SVM classifier, a five-fold cross-validation was implemented to minimize the bias generated by random sampling of the training and testing datasets. The input dataset was divided into five mutually exclusive subsets. One subset was used for testing and the remaining four were used to train the data. The process was repeated five times to ensure that the model was tested in each subset.

In Step 5, the most suitable SVM model was generated and compared with a logistic regression model to determine validity relative to our 54-item instrument. Initial data analysis was carried out by comparing the normalized importance, rank order of independent variables, and classification accuracy on the SVM to those of the logistic regression. The classification accuracy was compared by the measures of 1) sensitivity and 2) specificity, and 3) combined accuracy for each SVM classifier.

Sensitivity was defined as a measure of ability of the model to detect important variables associated with those respondents living in ZIP code's 37208, 37209 and 37210. Specificity is a measure of the ability of the model to specify and delineate differences by the respondents living in ZIP codes that make up the control group.

In Step 6, the most important variables and related rank orders of variables for the SVM model were generated to facilitate variable selection and explanation. The normalized importance was calculated by dividing the value of the highest relative importance into the value of the other relative importance. The normalized importance provides a hierarchal viewpoint of the ranking of the explanatory variables.

Creation of relational database and web portal for geographical information systems mapping

We have established a relational database and Web portal, in part, from the EPA databases listed at www.IMNashville.com. This Web portal enables us to 1) build the capacity to process and analyze large secondary databases; 2) conduct inter-disciplinary training; 3) use public participatory GIS; and 4) develop interactive mapping of health disparities and community assets and risk factors associated with environmental exposures. Emissions data from the EPA-databases listed below was converted into a mapping format to visually communicate the location of EPA reporting industrial plants with emissions in ZIP codes as a function of the ZIP codes of respondents to our 54-item instrument that perceived themselves vulnerable to exposure to contaminants.

USEPA databases used in the present study—The EPA reporting sites are defined as any site in Metro Nashville Davidson County that emissions data was available from the following EPA-databases 1) Air Quality Subsystem (AQS), 2) AIRS Facility Subsystem (AFS), 3) AIRS Data and 4) Toxic Release Inventory.

Air Quality Subsystem (AQS). The Air Quality System (AQS) is EPA's repository of ambient air quality data. AQS stores data from over 10,000 monitors, 5,000 of which are currently active. Specifically, this EPA database contains measurements of ambient concentrations of air pollutants and meteorological data from thousands of monitoring stations operated by EPA, state, and local agencies.

AIRS Facility Subsystem (AFS). The Air Facility System (AFS) contains compliance and enforcement data and permit data for stationary sources of air pollution regulated by EPA, state and local air pollution agencies. The environmental regulatory community uses this information to track the compliance status of point sources with various programs regulated under the Clean Air Act. See the Clean Air Act enforcement page for information on enforcement activities.

Types of data found in in AIRS facility subsystem—A *plant* is a facility represented by its physical location and defined by property boundaries. Plant-level data include plant name, address, Standard Industrial Classification (SIC), U.S. Census Bureau North American Industry Classification System, and compliance status.

A *stack* is where emissions are introduced into the atmosphere. Stack-level data include the height and diameter of the stack as well as the temperature, flow rate, and velocity of the gas released into the atmosphere. Stack-level data is used in emission inventory reporting.

A *point* is a physical piece of equipment or a process that produces emissions. Point-level data include normal operating schedule and the percentage of annual activity occurring each season.

A *segment* is a component of a point process, such as fuel combustion, that is used in the computation of emissions. Segment-level data in AFS include Source Classification Code (SCC), annual process rate, and fuel parameters. Segment-level data is used in emission

inventory reporting. Emission inventory data can be found at the Technology Transfer Network Clearinghouse.

AIRS data provide easy access to summaries of air monitoring data for the current and five prior years, the latest available estimates of air pollutant emissions from major point sources, the overall regulatory compliance status of those sources, and names of contacts in EPA and state/local air pollution agencies. All these data pertain to the criteria pollutants (carbon monoxide, nitrogen dioxide, sulfur dioxide, ozone, particulate matter, lead).

Toxics Release Inventory Program (TRI)—The TRI database contains data on disposal or other releases of over 650 toxic chemicals from thousands of U.S. facilities and information about how facilities manage those chemicals through recycling, energy recovery, and treatment. One of TRI's primary purposes is to inform communities about toxic chemical releases to the environment.

Results

Similarity of demographic characteristics between inner-city atlanta area (Grady Memorial Hospital) and inner-city nashville area (Metro Nashville General Hospital at Meharry Medical College)

A version of our 54-item instrument was administered previously to inner-city populations at Grady Memorial Hospital in Atlanta, Georgia. As a means of validating the instrument that was developed by our group, we sought to determine the similarity of demographics between inner-city Atlanta and inner-city Nashville based on 2000 and 2010 census data. For additional clarity, we also determined the similarity of demographics between the Grady Memorial Hospital area and the Metro Nashville General Hospital at Meharry area, as discussed in the Methods section and as demonstrated on Table 1. All of the endpoints and indices contributed to the validation -process of our 54-item instrument.

SVM analysis

Shown in Table 2 are the dependent or outcome variable as residential ZIP code areas that were regrouped into a binary measure of one or two. The respondent population was split between ZIP codes for the respondents living in (37206, 37207, 37208, 37209) and other ZIP codes for respondents living in (37210, 36606, 37011, 03713, 37062, 37075, 37076, 37082, 37086, 37087, 37115, 37219, 37138, 37203, 37210, 37211, 37212, 37213, 37214, 37215, 37216, 37217, 37218, 37219, 37220, 37221, 37228, 37235, and 38556). The five domains of independent variables (community, home and hobby, school, occupation, and environmental exposure history) included Q6 (polluted lake/stream: Community); Q17A (in your home: Home and Hobby); Q17C (on your lawn or garden: Home and Hobby); Q22 (school neighborhood [urban or rural]); Q24A (carpeted classrooms: School); Q24E (windows that open: School); Q24F (moldy smell: School); Q25G (flood, water leaks: School); and Q26G (odorous cleaning products: School).

Approximately one-tenth (9%) of the survey respondents lived near a polluted lake/stream community (Q6). In terms of using pesticides, more than half (54%) of the survey respondents used them in their home (Q17A) while more than one-tenth (14%) used them on

their lawn or garden (Q17C). More than half (53%) of the survey respondents indicated that their children's school was located in a rural area (Q22). While inspecting their children's schools, the survey respondents revealed that more than half of the schools had carpet (51%) (Q24A), windows that opened (55%) (Q24E), and a moldy smell (61%) (Q24F). Furthermore, nearly one-tenth (8%) of the survey respondents reported flood and/or water leaks in their children's school (Q25G), and more than three quarters (76%) of the survey respondents indicated that the schools used odorous cleaning products (Q26G).

Table 3 presents major findings using SVM and logistic regression modeling approaches with the latter approach being the traditional model used to measure a binary outcome. The backward procedure is one of the procedures used to select significant risk factors to form the logistic regression model. The model fitting the statistic, pseudo R-squared value was used to measure the success of this model by explaining the variations in the data. The Nagelkerke R squared value (0.32) was significantly different from zero, indicating that 32% of the variations in the binary outcome variable (residential ZIP codes) was accounted for by the risk factors. Using the SVM, the independent variables were assigned a normalized importance and given a rank order based on this normalized value. According to the sigmoid function of SVM, the variables by level of importance to the outcome were the following; Q17A (in your home), Q17C (on your lawn or garden), Q25G (flood, water leaks), Q6 (polluted lake/stream), Q24F (moldy smell), Q26G (odorous cleaning products), Q25E (child's school—new flooring or furniture), Q22 (school neighborhood [urban or rural]), and Q24A (carpeting in classrooms). The Wald test of Logistic regression was used at different levels of p values, 0.001, 0.01 or 0.05, to determine the significant independent variables associated with the binary outcome (residential ZIP codes). The regression coefficients for the following explanatory variables, Q17A, Q17C, Q25G, Q6, Q24F, Q26G, Q24A were found to be significantly different from zero at given p values level mentioned above and using the Wald test.

Table 4 represents a summary of the measures of accuracy used to evaluate both of the data analysis techniques used. Sixty-one percent (61%) of respondents perceived little to no exposure within the 5-domains. Sixty-nine percent (69%) of respondents perceived that they were susceptible to environmental exposures within the five domains. The SVM method outranked the logistic regression analysis with respect to sensitivity, but not specificity. In terms of combined accuracy, both SVM and logistic regression models were comparable (69% vs. 70%), which in some instances may be considered low; however, an important measure of accuracy is the sensitivity criterion, the so-called power of the test. Model sensitivity determines if the tool used in the study was accurate in measuring the outcome variable. If more survey respondents who were exposed to the risk factors did not have the disease, then the overall study power would be weakened. The SVM was found to be a more sensitive model (69%) than the logistic regression model (63%), meaning that SVM has a greater ability to identify important variables related to the binary outcome (residential ZIP codes). On the contrary, logistic regression was found to have a greater ability to specify differences between the populations of interest (women residing in the ZIP codes; 37206, 37207, 37208, 37209) and the control group used (women residing in other ZIP code areas).

Mapping of ePa-reporting industrial Sites located in the north nashville ZiP Code 37208. Figure 1 shows the EPA-Toxic Release Inventory (TRI) and smokestack emission sites located in the North Nashville ZIP code 37208. The TRI sites are shown in squares and smokestack emission sites in circles. Mapping of EPA-reporting industrial sites located in the North Nashville ZIP code 37208 allows for visualization of the SVM analysis finding of identifying the important variable related to the binary outcome as residential ZIP code from respondents of the 54-item instrument.

Discussion

In this study we validated our 54-item instrument with SVM analysis. The validation of this questionnaire was set in an inner city area of Nashville, Tennessee (North Nashville) which is composed largely of an African American population. North Nashville is known to suffer from increased adverse health outcomes compared with whites in the same area of Davidson County.³⁵ The respondent population at Grady Memorial Hospital in inner-city Atlanta and the respondent population at Metro Nashville General Hospital at Meharry were shown to be very similar in terms of various demographic characteristics including race, educational attainment, sex, and school enrollment. This 54-item instrument will be used in the very near future to recruit a large cohort of pregnant African American women into a longitudinal study designed to quantify perturbations in sensory function in their infants prior to 24 months. The primary study site, Metro Nashville General Hospital on the Campus of Meharry Medical College operates as the community's public hospital to provide access to care for all the city's residents—whether uninsured, underinsured or insured. As a result of the validation of our instrument, our hospital will continue to serve as the primary recruitment site for translational research projects under the Meharry-Vanderbilt Alliance umbrella. It will be an overarching hypothesis in the future that women living in ZIP codes 37208 and 37209 will have a higher perceived risk of exposure to environmental toxicants as compared to women living in the surrounding suburban areas.

Support vector machines have been widely used in various areas, such as recognition, reliability evaluation, bioinformatics and medicines, for survival time classification and assessment of the severity of many acute and chronic diseases.³⁶ However, to our knowledge, this study is the first attempt to investigate perceived exposure to environmental contaminants by administration of a 54-item instrument in a population of child-bearing-age African American and Hispanic women *via* SVM. The novelty of our study was demonstrating the discriminative power of this SVM approach as compared with that of commonly used logistic regression models.

In association with other studies being conducted in our center, creation of environmental relational databases will serve as a means to promote civic engagement in vulnerable communities and neighborhoods in close proximity to Meharry Medical College. The database described here enables us to begin to analyze the relationships between physical, built, and social environments relative to the health status of persons living in targeted neighborhoods or communities. The long-term goal is to use the data to develop and implement targeted interventions to improve health conditions on a community level. Our Web portal (www.IMNashville.com) allowed us to bring together data on health outcomes

with data on environmental conditions, assets (salutogenic features), and risk factors (pathogenic features). As indicated earlier, our approach builds on the ecological environmental justice framework that explores the role that salutogenic (health-promoting) and pathogenic (health-restricting) built and social features play in driving health and health disparities in differentially burdened communities.³⁷

For the purposes of such relational databases, the natural environment includes air, food, soil, water, and vegetation.³⁸ Environmental hazards refer to the effects of the environment on a person's health and the effects of human activity on the health of the environment. Such hazards can be toxicants in a physical space that can harm health. Natural environmental datasets in our relational database include: 1) Day and night land surface temperature (LST) data from NASA's Moderate-Resolution Imaging Spectroradiometer (MODIS) satellite sensor (1-km spatial resolution); 2) Daily spatial surfaces of ambient fine particulate matter (PM_{2.5}) generated by algorithms developed by scientists at the Universities Space Research Association (USRA) and NASA/Marshall Space Flight Center and using MODIS satellite data and EPA ground observations³⁹ (10-km resolution); 3) Daily maximum/minimum air temperature and daily maximum heat index from the North American Land Data Assimilation System (NLDAS) forcing data (12-km resolution); and 4) Finer resolution (30-m and 60-m) datasets of Landsat-derived land cover/land use and LST for 1992 and 2006 for a focused study in Nashville/Davidson County.

The built environment has been found to play a major role in directing individual physical activity, and Nashville has made big strides in improving policies and regulations related to building and site design to improve the built environment for pedestrians and cyclists, including passage of (1) specific plan zoning; (2) revised subdivision regulations that have introduced a so-called *walkable subdivision* option for developers; and (3) a community-character manual that will guide future land-use planning.⁴⁰ In the future, our relational database will allow for the analysis of relationships between changing characteristics of the built environment and occurrence of health disparities. Establishment of this relational database will allow us to build longitudinal datasets to track changes in environmental conditions and effects on health disparities both retrospectively and prospectively.

Our study however, is not without limitations. The primary limitation of this study is its relatively small sample size but this fact did not impact the ability to achieve statistical significance with regard to the determination of higher ordered relationships to responses on the 54-item instrument. The second limitation is that not all of the women who participated were pregnant at the time of interview. We are aware of the implications of this fact, and remain cautious as a result. However, given the demographic profile of the patient base at Metro Nashville General Hospital at Meharry Medical College, it is highly likely that the percent of unintended births will track with that observed in the United States (37%).⁴¹ Many studies on unintended childbearing including a comprehensive review by the Institute of Medicine⁴² and a recent white paper reviewing more than 60 additional studies on this topic,⁴³ have also shown that births that were unintended by the mother are at elevated risk of adverse social, economic, and health outcomes for the mother and the child. Unintended births are associated with delayed prenatal care, smoking during pregnancy, not breastfeeding the baby, poorer health during childhood, and poorer outcomes for the mother

and the mother-child relationship.⁴⁴ Further, longer-term negative consequences for these children have been found by some longitudinal studies of unintended pregnancies that track the children into adulthood.⁴⁵

Thirdly, our cross-sectional design does not allow any inference to be drawn with respect to the causal relationships among independent variables. Finally, our data is based on a 15-minute interview in an indigent care facility that is located in a known socioeconomically impoverished ZIP code. This fact may have contributed to the perception of exposure to pollutants as has been observed in Berkson's bias.⁴⁶ Connecting environmental problems with health disparities can be difficult due to limitations in the way health data are collected and made available.⁴⁷ Many health databases and tables provide data only at the county level. Without a geo-coded reference at a sub-county level, these data are limited as to how they can be related to environmental hazards or exposures that typically are features of a small geographic area. In addition, the health data often tells us very little about how a person may have come in contact with an environmental hazard.

Conclusion

To our knowledge, this is the first study using SVM analysis determination of higher-ordered spatial relationships on an environmental exposure history questionnaire. Previous to this one, the study that most closely resembled this one used SVM to predict adherence to medication in heart failure patients. In that study, data about medication adherence was collected from patients at a university hospital through self-reported questionnaire. The data included 11 variables of 76 patients with heart failure. Mathematical simulations were conducted in order to develop a SVM model for the identification of variables that would best predict medication adherence. Using the sigmoid function via SVM analysis,⁴⁷ the rank order of variables by level of importance to the outcome (of perceived exposure to environmental contaminants) were Q17A (in your home), Q17C (on your lawn or garden), Q25G (flood, water leaks), Q6 (polluted lake/stream), Q24F (moldy smell), Q26G (odorous cleaning products), Q25E (child's school—new flooring or furniture), Q22 (school neighborhood—urban or rural), and Q24A (carpeting classrooms). The Wald test of logistic regression was used to determine the significance (at different levels of p values: .001, .01, or .05) of independent variables associated with the binary outcome (residential ZIP codes). The regression coefficients for Q17A, Q17C, Q25G, Q6, Q24F, Q26G, Q24A were found to be significantly different from zero at given p value level. Our Nagelkerke R squared value (0.32) was significantly different from zero, indicating that 32% of the variations in the binary outcome variable (residential ZIP codes) were accounted for by the risk factors indicated in the respective questions. In conclusion, SVM analysis was shown to have discriminate power for determination of higher-ordered spatial relationships on the 54-item instrument within the five domains studied in the target ZIP code.

Acknowledgments

This work was supported, in part, by NIH grants S11ES014156-06 and 1R56ES017448-01A1 to DBH and 3P20MD000516-07S1 to PDJ and DBH. Also, critical to the conduct of these studies were grants from the Simons Foundation Autism Research Initiative, Research Centers in Minority Institutions (RCMI) G12RR03032 and S06GM08037, Nuclear Regulatory Commission Grant NRC-27-10-515 as well as Meharry Medical College - Vanderbilt University Alliance for Research Training in Neuroscience Grant (T32MH065782).

Notes

1. Raymond J, Wheeler W, Brown MJ, et al. Inadequate and unhealthy Housing, 2007 and 2009. *MMWR*. Jan; 2011 60(Suppl):21–27. [PubMed: 21430615]
2. Berube, A.; Frey, WH.; Singer, A., et al. Getting current: recent demographic trends in metropolitan America. Brookings Institution; Washington DC: 2009. Available at: <http://www.brookings.edu/research/reports/2009/03/metro-demographic-trends>
3. U.S. Environmental Protection Agency. Integrated science assessment for particulate matter : final report. U.S. Environmental Protection Agency; Washington, DC: 2009. Available at: <http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid5216546>
4. U.S. Environmental Protection Agency. Air quality criteria for ozone and related photochemical oxidants: final report. U.S. Environmental Protection Agency; Washington, DC: 2006. Available at: <http://cfpub.epa.gov/ncea/cfm/recordisplay.cfm?deid5149923>
5. Cirera L, Rodriguez M, Gimenez J, et al. Effects of public health inter ventions on industrial emissions and ambient air in Cartagena, Spain. *Environ Sci Pollut Res Int*. Mar; 2009 16(2):152–61. [PubMed: 19104864]
6. Johansson C, Burman L, Forsberg B. The effects of congestions tax on air quality and health. *Atmospheric Environment*. Oct; 2009 43(31):4843–54.
7. Pope CA 3rd. Respiratory hospital admissions associated with PM10 pollution in Utah, Salt Lake, and Cache Valleys. *Arch Environ Health*. Mar-Apr;1991 46(2):90–7. [PubMed: 2006899]
8. Wang S, Zhao M, Xing J, et al. Quantifying the air pollutants emission reduction during the 2008 Olympic games in Beijing. *Environ Sci Technol*. Apr; 2010 44(7):2490–6. [PubMed: 20222727]
9. Currie J, Neidell M, Schmieder JF. Air pollution and infant health: lessons from New Jersey. *J Health Econ*. May; 2009 28(3):688–703. [PubMed: 19328569]
10. Ritz B, Yu F. The effect of ambient carbon monoxide on low birth weight among children born in southern California between 1989 and 1993. *Environ Health Perspect*. Jan; 1999 107(1):17–25. [PubMed: 9872713]
11. Hansen CA, Barnett AG, Pritchard G. The effect of ambient air pollution during early pregnancy on fetal ultrasonic measurements during mid-pregnancy. *Environ Health Perspect*. Mar; 2008 116(3):362–9. [PubMed: 18335104]
12. Vassilev ZP, Robson MG, Klotz JB. Outdoor exposure to airborne polycyclic organic matter and adverse reproductive outcomes: a pilot study. *Am J Ind Med*. Sep; 2001 40(3):255–62. [PubMed: 11598971]
13. Bocskay KA, Tang D, Orjuela MA, et al. Chromosomal aberrations in cord blood are associated with prenatal exposure to carcinogenic polycyclic aromatic hydrocarbons. *Cancer Epidemiol Biomarkers Prev*. Feb; 2005 14(2):506–11. [PubMed: 15734979]
14. Perera FP, Rauh V, Whyatt RM, et al. Molecular evidence of an interaction between prenatal environmental exposures and birth outcomes in a multiethnic population. *Environ Health Perspect*. Apr; 2004 112(5):626–30. [PubMed: 15064172]
15. Perera FP, Tang D, Rauh V, et al. Relationship between polycyclic aromatic hydro-carbon-DNA adducts, environmental tobacco smoke, and child development in the World Trade Center cohort. *Environ Health Perspect*. Oct; 2007 115(10):1497–502. [PubMed: 17938742]
16. Perera FP, Rauh V, Whyatt RM, et al. Effect of prenatal exposure to airborne polycyclic aromatic hydrocarbons on neurodevelopment in the first 3 years of life among inner-city children. *Environ Health Perspect*. Aug; 2006 114(8):1287–92. [PubMed: 16882541]
17. Perera FP, Rauh V, Tsai WY, et al. Effects of transplacental exposure to environmental pollutants on birth outcomes in a multiethnic population. *Environ Health Perspect*. Feb; 2003 111(2):201–5. [PubMed: 12573906]
18. Morales E, Julvez J, Torrent M, et al. Association of early-life exposure to household gas appliances and indoor nitrogen dioxide with cognition and attention behavior in preschoolers. *Am J Epidemiol*. Jun; 2009 169(11):1327–36. [PubMed: 19395695]
19. Suglia SF, Gryparis A, Wright RO, et al. Association of black carbon with cognition among children in a prospective birth cohort study. *Am J Epidemiol*. Feb; 2008 167(3):280–6. [PubMed: 18006900]

20. Windham GC, Zhang L, Gunier R, et al. Autism spectrum disorders in relation to distribution of hazardous air pollutants in the San Francisco Bay Area. *Environ Health Perspect.* Sep; 2006 114(9):1438–44. [PubMed: 16966102]
21. Cortes, C.; Vapnik, V. Kluwer Academic Publishers/Machine Learning; Boston, MA: 1995. Support-vector networks.; p. 273-97. Available at: http://image.diku.dk/imagecanon/material/cortes_vapnik95.pdf
22. Verplancke T, Van Looy S, et al. Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haemato-logical malignancies. *BMC Med Inform Decis Mak.* Dec 5.2008 8:56. [PubMed: 19061509]
23. Yu W, Liu T, Valdez R, et al. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and prediabetes. *BMC Med Inform Decis Mak.* Mar 22.2010 10:16. [PubMed: 20307319]
24. Maglogiannis I, Loukis E, Zafiropoulos E, et al. Support vectors machine-based identification of heart valve diseases using heart sounds. *Comput Methods Programs Biomed.* Jul; 2009 95(1):47–61. [PubMed: 19269056]
25. Thurston RC, Matthews KA, Hernandez J, et al. Improving the performance of physiologic hot flash measures with support vector machines. *Psychophysiology.* Mar; 2009 46(2):285–92. [PubMed: 19170952]
26. Agency for Toxic Substances and Disease Registry (ATSDR). Case studies in environmental medicine: taking an exposure history. Agency for Toxic Substances and Disease Registry; Washington, DC: Oct. 1992 Available at: http://www.atsdr.cdc.gov/hec/csem/exphistory/docs/exposure_history.pdf
27. Rice SB, Nenadic G, Stapley BJ. Mining protein function from text using term-based support vector machines. *BMC Bioinformatics.* 2005; 6(Suppl 1):S22. [PubMed: 15960835]
28. Ng KL, Mishra SK. De Novo SVM Classification of Precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics.* Jun 1; 2007 23(11):1321–30. [PubMed: 17267435]
29. Vapnik, V. The nature of statistical learning theory and application. John Wiley Publishing; New York, NY: 1995.
30. Wan, V.; Campbell, WM. Support vector machines for speaker verification and identification.. Presented at: IEEE International Workshop on Neural Networks for Signal Processing; Sydney, Australia. 2000.
31. Zeng, D.; Xu, J.; Gu, J., et al. Short term traffic flow prediction based on online learning SVR. IEEE Computer Society/Proceedings of the Workshop on Power Electronics and Intelligent Transportation System; Washington, DC: 2008. p. 616-20.
32. Vapnik, V. The Nature of statistical learning theory. 2nd ed.. Springer-Verlag; New York, NY: 1999.
33. Cover, TM. Geometrical and statistical properties of system of linear inequalities with applications in the pattern recognition.; IEEE Transaction on Electronic Computers. 1965. p. 326-34. Available at: <http://hebb.mit.edu/courses/9.641/2002/readings/Cover65.pdf>
34. Sherrod, PH. DTREG Software; Brentwood, TN: 2010. Digital tree regression software for predictive modeling and forecasting (Version 4.0).. Available at: <http://www.dtreg.com>
35. Blaha MJ, Kusz KL, Drake W, et al. Hypertension prevalence awareness, treatment and control in North Nashville. *Tenn Med.* Apr; 2006 99(4):35–7. [PubMed: 16704130]
36. Cho BH, Yu H, Kim KW, et al. Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. *Artif Intell Med.* Jan; 2008 42(1): 37–53. [PubMed: 17997291]
37. Wilson SM, Wilson OR, Heaney CD, et al. Use of EPA collaborative problem-solving model to obtain environmental justice in North Carolina. *Prog Community Health Partnersh.* 2007; 1(4): 327–37. Winter. [PubMed: 20208212]
38. Ramesh, A.; Archibong, A.; Hood, DB., et al. Global environmental distribution and human health effects of polycyclic aromatic hydrocarbons.. In: Loganathan, BG.; Lam, PK., editors. Global trends of persistent organic chemicals. Boca Raton, FL: 2012. p. 97-126.

39. Al-Hamdan MZ, Crosson WL, Limaye AS, et al. Methods for characterizing fine particulate matter using ground observations and remotely sensed data: potential use for environmental public health surveillance. *J Air Waste Manag Assoc.* Jul; 2009 59(7):865–81. [PubMed: 19645271]
40. Omishakin AA, Carlat JL, Hornsby S, et al. Achieving built-environment and active living goals through Music City moves. *Am J Prev Med.* Dec; 2009 37(6 Suppl 2):S412–9. [PubMed: 19944942]
41. Mosher WD, Jones J, Abma JC. Intended and unintended births in the United States: 1982–2010. *Natl Health Stat Report.* Jul 24.2012 (55):1–28.
42. Brown, SS.; Eisenberg, L. *The best intentions: unintended pregnancy and the well-being of children and families.* National Academy Press; Washington, DC: 1995.
43. Logan, C.; Holcombe, E.; Manlove, J., et al. *The consequences of unintended childbearing: a white paper.* The National Campaign to Prevent Teen and Unplanned Pregnancy; Washington, DC: 2007. available at: <http://www.thenationalcampaign.org/resources/pdf/consequences.pdf>
44. Barber JS, Axinn WG, Thornton A. Unwanted childbearing, health, and mother-child relationships. *J Health Soc Behav.* Sep; 1999 40(3):231–57. [PubMed: 10513146]
45. David HP. Born unwanted, 35 years later: The Prague Study. *Reprod Health Matters.* May; 2006 14(27):181–90. [PubMed: 16713893]
46. Berkson J. Limitations of the application of fourfold table analysis to hospital data. *Biometrics.* Jun; 1946 2(3):47–53. [PubMed: 21001024]
47. Long J, Li Y, Yu Z. A semi-supervised support vector machine approach for parameter setting in motor imagery-based brain computer interfaces. *Cogn Neurodyn.* Sep; 2010 4(3):207–16. [PubMed: 21886673]

Figure 1.

Mapping of EPA-Toxic Release Inventory and smokestack emission sites located in the North Nashville zip code 37208 coincides with higher perceived exposure to environmental contaminants by respondents. Data from our NIMHD-sponsored Health Disparities Relational Database was used to map EPA Toxic Release Inventory (TRI) and smokestack emission sites within Metro Nashville Davidson County. The TRI sites are shown in squares and smokestack emission sites in circles for the zip code 37208.

Table 1

DEMONSTRATION OF SIMILARITY OF DEMOGRAPHICS BETWEEN INNER-CITY ATLANTA AREA (GRADY MEMORIAL HOSPITAL) AND INNER-CITY NASHVILLE AREA (METRO NASHVILLE GENERAL HOSPITAL)

Subject	Atlanta, Georgia 2010 Census (Atlanta City-Fulton)	Grady Memorial Hospital Area 2000 Census (Zip Code 30303)	Nashville, Tennessee 2010 Census (Nashville-Davidson)	Metro Nashville General Hospital Area 2000 Census (Zip Code 37208)	Atlanta, GA/ Nashville-Davidson 2010 Census	Grady Hospital/ Metro General Hospital 2000 Census	Percent Difference
Total Population	920,581 (100%)	3,744 (100%)	626,681 (100%)	15,272 (100%)	—	—	—
Sex							
Male	448,267 (48.7%)	2,631 (70.3%)	303,540 (48.4%)	6,994 (45.8%)	.3		24.5
Female	472,314 (51.3%)	1,113 (29.7%)	323,141 (51.6%)	8,278 (54.2%)	-.3		-24.5
Age (years)							
15-34	287,560 (31.2%)	1,665 (45.1%)	206,879 (33.0%)	4,549 (29.8%)	-1.8		15.3
35 and above	449,045 (48.8%)	1,786 (47.7%)	303,594 (48.4%)	7,306 (47.8%)	.4		-0.1
School Enrollment in 2000							
Population (3 years & over) enrolled in school	219,663 (100%)	780 (100%)	141,200 (100%)	5,053 (100%)	—	—	—
Less than high school	122,441 (55.7%)	294 (37.7%)	72,863 (51.6%)	2,520 (49.9%)	4.1		-12.2
High school	42,144 (19.2%)	221 (28.3%)	27,016 (19.1%)	1,054 (20.9%)	.1		7.4
College and above	55,078 (25.1%)	265 (34.0%)	41,321 (29.3%)	1,479 (29.3%)	-4.2		4.7
Population (25 years and over)	527,738 (100%)	2,969 (100%)	377,734 (100%)	9,222 (100%)	—	—	—
Less than high school	84,370 (16%)	1,158 (39.0%)	69,801 (18.5%)	3,522 (38.1%)	-2.5		.9
High school graduate	102,246 (19.4%)	772 (26.0%)	93,105 (24.6%)	2,472 (26.8%)	-5.2		-8
College and above	341,122 (64.6%)	1,039 (35.0%)	214,828 (56.9%)	3,228 (35.0%)	7.7		.0
Race							
White	409,697 (44.5%)	700 (18.7%)	385,039 (61.4%)	703 (4.6%)	-16.9		14.1
Black or African American	405,575 (44.1%)	2,857 (76.3%)	173,730 (27.7%)	14,162 (92.7%)	16.4		-16.4
American Indian and Alaska Native	2,259 (0.2%)	15 (0.4%)	2,091 (0.33%)	39 (0.3%)	-1.3		.1
Asian	51,569 (5.6%)	39 (1.0%)	19,027 (3.0%)	68 (0.4%)	2.6		.6

Subject	Atlanta, Georgia 2010 Census (Atlanta City- Fulton)	Grady Memorial Hospital Area 2000 Census (Zip Code 30303)	Nashville, Tennessee 2010 Census (Nashville- Davidson)	Metro Nashville General Hospital Area 2000 Census (Zip Code 37208)	Percent Difference	
	Number (Percent)	Number (Percent)	Number (Percent)	Number (Percent)	Atlanta, GA/ Nashville- Davidson 2010 Census	Grady Hospital/ Metro General Hospital 2000 Census
Native Hawaiian and other Pacific Islander	335 (<0.1%)	6 (0.2%)	394 (<0.1%)	16 (0.1%)	0	.1

Table 2**STUDY VARIABLES**

Variable Names	Variable Descriptions
Dependent Variable	
ZipGroup	1–aggregation of zip code areas: 37206, 37207, 37208 and 37209 and 2–aggregation of zip code areas: 37210, 36606, 37011, 37013, 37062, 37075, 37076, 37082, 37086, 37087, 37115, 37219, 37138, 37203, 37210, 37211, 37212, 37213, 37214, 37215, 37216, 37217, 37218, 37219, 37220, 37221, 37228, 37235, and 38556
Independent Variables	
Q6	Polluted lake/stream (yes/no)
Q17A	In your home (yes/no)
Q17C	On your lawn or garden (yes/no)
Q24A	Carpeted classrooms
Q24E	Windows that open
Q24F	Moldy smell
Q25E	Child's school–new flooring or furniture
Q25G	Flood, water leaks
Q26G	Odorous cleaning products
Q22	School neighborhood (rural/urban)

Table 3**SVM AND LOGISTIC REGRESSION MODELS**

Variable in Equation	Model A–SVM Model: (Sigmoid) ^a		Model B–logistic regression Model: backward (wald) ^b	
	Normalized importance	Rank Order	Regression Coefficient	P Value
Q17A	1.00	1	.963	.012 ^c
Q17C	.67	2	−1.478	.010 ^c
Q25G	.40	3	−2.965	.007 ^d
Q6	.27	4	−1.522	.016 ^c
Q24F	.20	5	1.082	.024 ^c
Q26G	.07	6	−2.128	.0001 ^e
Q25E	Close to 0	7	.767	.067
Q22	Close to 0	8	.807	.088
Q24A	Close to 0	9	−.981	.039 ^c

SVM = Support Vector Machine

^a Sigmoid kernel function with C = 50000; Gamma = 0.003 and constant term = 0.203

^b The model fitting statistic for the logistic regression model is Nagelkerke R squared value of 0.32.

^c Significant variables with p<0.05 using the Wald test in the logistic regression model

^d Significant variables with p<0.01 using the Wald test in the logistic regression model

^e Significant variables with p<0.001 using the Wald test in the logistic regression model

Table 4

COMPARISON OF CLASSIFICATION

Criteria for Accuracy	Model A SVM Model: Sigmoid	Model B Logistic Regression Model: Backward (Wald)
Combined Accuracy	69%	70%
Sensitivity	69%	63%
Specificity	61%	69%

SVM = Support Vector Machine