Spatial and Spatio-temporal Epidemiology 21 (2017) 13–23

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

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

Original Research

Exploring racial disparity in obesity: A mediation analysis considering geo-coded environmental factors

Qingzhao Yu[a,*], Richard A. Scribner[b], Claudia Leonardi[b], Lu Zhang[c], Chi Park[b], Liwei Chen[c], Neal R. Simonsen[d]

[a] Biostatistics Program, School of Public Health, Louisiana State University Health Sciences Center, 3rd floor, 2020 Gravier Street, New Orleans, LA 70112, United States
[b] Epidemiology Program, School of Public Health, Louisiana State University Health Sciences Center, New Orleans, LA, United States
[c] Department of Public Health Sciences, Clemson University, Clemson, SC, United States
[d] New Orleans, LA, United States

ARTICLE INFO

Article history:
Received 5 December 2016
Revised 31 January 2017
Accepted 10 February 2017
Available online 17 February 2017

Keywords:
Mediation analysis
Obesity
Physical activity
Racial disparity

ABSTRACT

Research shows a consistent racial disparity in obesity between white and black adults in the United States. Accounting for the disparity is a challenge given the variety of the contributing factors, the nature of the association, and the multilevel relationships among the factors. We used the multivariable mediation analysis (MMA) method to explore the racial disparity in obesity considering not only the individual behavior but also geospatially derived environmental risk factors. Results from generalized linear models (GLM) were compared with those from multiple additive regression trees (MART) which allow for hierarchical data structure, and fitting of nonlinear and complex interactive relationships. As results, both individual and geographically defined factors contributed to the racial disparity in obesity. MART performed better than GLM models in that MART explained a larger proportion of the racial disparity in obesity. However, there remained disparities that cannot be explained by factors collected in this study.

1. Introduction

Obesity is a serious public health concern in the United States (US). More than one third of US adults have obesity. It is well documented that black people have a higher rate of obesity than white people, which results in the racial disparity in obesity and the related diseases. Explanations for the obesity epidemic and the respective racial disparity are multifactorial. In addition to the traditional individual behavior risk factors, geo-coded risk factors at the neighborhood level have also been proposed. Aspects of the built environment have been considered as modifiable risk factors that influence both energy expenditure and energy consumption and therefore can be modified to address the obesity epidemic as well as the disparity in obesity between blacks and whites (Hill et al., 2003; Khan et al., 2009; Papas et al., 2007; Roux, 2003). Efforts to

Abbreviations: US, United States; NHANES, National Health and Nutrition Examination Survey; GIS, Geographic Information System; MART, multivariate additive regression trees; NCHS, National Center for Health Statistics; CDC, the Centers for Disease Control and Prevention; MEC, Medical Examination Component; NCAIS, North American Industry Classification System; SIC, Standard Industrial Classification; ESRI, Environmental Systems Research Institute; GLM, Generalized Linear Model; BMI, body mass index; kg, kilograms; m^2, squared meters; ANOVA, analysis of variance; GLM, generalized linear models; CDI, Concentrated Disadvantage Index.

* Corresponding author.
E-mail addresses: qyu@lsuhsc.edu (Q. Yu), rscrib@lsuhsc.edu (R.A. Scribner), cleon1@lsuhsc.edu (C. Leonardi), lz3@clemson.edu (L. Zhang), liz0817@gmail.com (C. Park), liweic@clemson.edu (L. Chen), epiman@yahoo.com (N.R. Simonsen).

http://dx.doi.org/10.1016/j.sste.2017.02.001
1877-5845/

Table 1
Measures of street connectivity used in the analyses.

Measure (notation)	Definition	Calculation*
Intersection density	Number of intersections per square mile of area	# Real nodes/area
Street density	Number of linear miles of street per square mile of area	Total # roadway miles/area
Connected node ratio	Ratio of number of intersections with four or more connections over the total number intersections	# of intersections associated with four or more links/total # of intersections

* Calculations utilize geographic information system-derived data.

understand the relative contribution of any single factor often result in equivocal findings. As a result, more sophisticated methods capable of addressing the variety and hierarchical structures are needed. Specifically, methods accounting for the contributions of both individual and environmental factors as well as the complex relationships (e.g., nonlinear) potentially involved. There has been considerable interest in creating spatially defined risk factors, and quantifying these risk factors at the neighborhood level in order to identify modifiable factors in the environment that account for obesity in general and the disparity in particular (Ding and Gebel, 2012; Gebel et al., 2007; Grasser et al., 2013).

Walkability represents one of the factors that has a complex relationship with both obesity and the racial disparity in obesity rates among Americans. We created the variables: intersection density, street density, and connection node ratio (defined in Table 1, geospatial definitions are provided in the Method section) to measure the overall construct of neighborhood walkability. While the overall neighborhood walkability construct is widely considered as a factor associated with increased physical activity and lower rates of obesity, the research linking measures of the sub-construct street connectivity to obesity in general (Ball et al., 2012; Heinrich et al., 2008; Li et al., 2008; McDonald et al., 2012; Wang et al., 2013; Wen and Kowaleski-Jones, 2012), and physical activity in particular (Berrigan and Troiano, 2002; Eriksson et al., 2012; Frank et al., 2008; Li et al., 2005; Oakes et al., 2007; Owen et al., 2007; Pearce and Maddison, 2011; Saelens et al., 2003; Witten et al., 2012) has been equivocal, evidenced in the conclusions of some reviews on the topic (Grasser et al., 2013; Saelens and Handy, 2008). Street connectivity (e.g. intersection density) appears to be associated with both obesity and physical activity. However, when the effect of other measures of walkability (e.g., street density) are controlled, the association is attenuated. It is not clear whether the effect of street connectivity is explained by these other constructs or whether other statistical issues are involved. For example, the effect of street connectivity may have a nonlinear or even non-monotone relationship with the prevalence of obesity, leading to the equivocal results. Alternatively, the effect of street connectivity may be explained in terms of its' strong association with street density, which is actually driving the effect and apparent street connectivity effects arise from this multicollinearity. In any case, the relationship is complex and hierarchical. It poses serious problems for jointly using individual and geo-spatial factors in explanation in terms of existing analytic designs.

With regard to the disparity in obesity between whites and blacks, street connectivity also has a complex relationship. While blacks in general are more obese than whites, they are more likely to reside in urban centers. However, urban centers tend to be areas with high street connectivity suggesting that blacks living in these areas should be less obese than their white counterparts in low street connectivity areas (e.g., suburbs). In addition, King (2013) recently concluded that the increased density of urban centers may overcome the benefits of walking. Once again the relationship is a complex one involving a variety of factors that are difficult to characterize with existing analytic models.

Mediation effect refers to the effect conveyed by an intervening variable to an observed relationship between an exposure and a response variable of interest. To explore the racial disparity in obesity, mediation analysis is used where race is the exposure variable and obesity or its absence is the binary response. All other risk factors that may explain the racial disparity are considered as potential mediators. There are several key challenges in the exploration of racial disparity in obesity. First, the model should be able to deal with different types of mediators, where the potential factors can be continuous, binary or categorical with or without order. Second, the indirect effect from each mediator should be differentiable so the indirect effect conveyed by different factors can be compared. Third, given multiple measures of walkability with influences that might vary by measure, their mediating effect on the racial disparity needs to be assessed jointly rather than additively, so we need a method that can measure the joint effects from combined factors. Finally, there are potential nonlinear relationships and interactions among race, the mediators, and the obesity risk, therefore the fitted model should be able to represent the most reasonable adequately complex one that represents the relationships existing among variables.

There are generally two settings for mediation analysis. One is based on linear models (Baron and Kenny, 1986; MacKinnon et al., 1995), and the other is based on the counterfactual framework (Albert, 2008; Have et al., 2007; Pearl, 2001; Robins and Greenland, 1992). In this paper, we adapted a general definition of mediation effect by Yu et al. (2014) based on the counterfactual framework. The derived mediation analysis proposed by Yu et al. is promising in that the indirect effects contributed by different mediators are separable, which enables the comparison of relative mediation effects carried by different third variables. The mediation analysis is generalized so that we can deal with binary, multi-categorical or continuous exposure, mediator and response variables. Moreover, general predictive models, as well as general linear models can be used

to fit variable relationships. The R package, mma (Yu and Li, 2017), was used for the data analysis in this study.

To assess the utility of this type of mediation analysis for complex relationships like that among race, risk factors from both individual and environmental levels, and obesity, we first created environmental risk factors at census tract level, and then linked the factors with the geocoded National Health and Nutrition Examination Survey (NHANES). The linkage allows for the characterization of both neighborhood and individual measures that take into account the variety of factors linked to this relationship so that the racial disparity in obesity can be explored. The statistical method can handle multiple mediators of different types and complex hierarchical inter-relationships, and the method can differentiate the indirect effects from each individual factor or joint factors. The rest of the paper is organized such that Section 2 introduces the NHANES and the generation of environment data sets using Geographic Information System (GIS). We discuss the methods of analysis and the rationale for their use. Details of the algorithms for model fitting and interpretations are discussed. Section 3 presents the results of the analysis, which includes inferences on the indirect effects from individual and joint factors that contribute to the racial disparity in obesity, and the dependence and interaction figures that describe the relationship among variables. The results from logistic regression and multivariate additive regression trees (MART) are compared and discussed. A discussion and conclusions regarding analytical results and the comparison of methods are presented in Section 4.

2. Methods

2.1. Data sets used in the study

This study linked data from the NHANES with neighborhood and census tract level sociodemographic and spatial data. In this section, we first introduce the combined dataset and then describe the measures and corresponding variables used in this study.

2.1.1. NHANES

The NHANES survey is a cross-sectional study conducted by the National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention (CDC) from 1971 onward, with biennial surveys beginning in 1999 ("Continuous NHANES"). The study is conducted through a national multistage probability sample of the civilian non-institutionalized population of the United States age 2 months and over (NCHS 1994). The survey captures an array of health information on various physiologic measures, health outcomes and diseases, and health behaviors through three formats: (1) survey, (2) medical exam, and (3) laboratory test. The survey design ensured oversampling of non-Hispanic blacks, Hispanics, persons 2 months to 5 years of age, and those $\geq$ 60 years of age. The sampling scheme for NHANES has remained fairly consistent over time, with the Primary Sampling Units generally single counties and clusters of households in each county selected. Each person in a selected household is screened for demographic characteristics, and one or more persons per household are selected for the sample. Data were collected from each participant through a face-to-face household interview. In addition, participants were invited to take a physical examination and provide biospecimens in the Medical Examination Component (MEC).

We utilized the Continuous NHANES 2003–2006 data for this study. The average sample size for biennial Continuous NHANES surveys is approximately 10,000 individuals age 2 months and older. This analysis was restricted to the non-pregnant adult (age $\geq$ 20) participants. Existing NHANES data were linked to separately collected neighborhood environmental data. Selected surveys from the NHANES, both housed and conducted through the National Center for Health Statistics (NCHS), have been georeferenced to improve the utility of data for research purposes (Harris, 2006).

2.2.2. Census tract level contextual data

Neighborhood and community level data were linked to the survey participants through their geographic identifiers. A number of data sources were accessed to generate measures of the neighborhood environment in general and the neighborhood food environment in particular.

The sources of the neighborhood data are the sociodemographic data from US Census 2000 and the American Community Survey, geographic data from ArcGIS and US Census shapefiles, and food environment information drawn from North American Industry Classification System (NCAIS), Standard Industrial Classification (SIC) data from InfoUSA and Environmental Systems Research Institute (ESRI). NCAIS/SIC data obtained include business name, geocoded location, and detailed SIC industry codes for food establishments. Census data include various measures as described below.

Given that the temporal frame of the georeferenced Continuous NHANES data ranges from 1999 through 2010, the project temporally aligned the data collected from different sources. Census derived measures were drawn from the 2000 and 2010 census and linearly interpolated between decennial censuses. NCAIS data were obtained historically at five year intervals going back to 2000. Estimates for measures between interval points were also linearly interpolated, where possible, informed by the American Community Survey. The exception was the Census shape files. All definitions of neighborhood were based on the Census 2000 Topologically Integrated Geographic Encoding and Referencing shape file.

2.2. Measures

2.2.1. Individual level measures

Key measures used from NHANES are categorized below.

Obesity related impact variables: the primary impact measures for the study were body mass index (BMI), defined as weight in kilograms (kg) divided by height in squared meters (m^2), and obesity, defined as having a BMI $\geq$ 30. Analyses were conducted for obesity status as a binary variable.

Dietary behavior variables: individual dietary variables in NHANES are obtained from a 24-h dietary recall, as well as a questionnaire on dietary behavior. Specifically, two key dietary factors typically associated with obesity were examined in this study, (1) total energy intake and (2) sugar sweetened beverage consumption. Both variables were operationalized as tertiles characterizing low, medium, and high categories.

Physical activity variable: physical activity was assessed using the accelerometry data available in the 2003–2004 and 2005–2006 cycles of the NHANES. The primary variables of interest were total energy expenditure and the level of physical activity—low to none, light, moderate, vigorous—defined based on the Metabolic Equivalent Task method. The physical activity variable was dichotomized as none to light physical activity and moderate to vigorous physical activity.

Control variables: other covariates considered in the analysis at the individual level included age, sex, race/ethnicity, education, family history of disease, language used/spoken at home, type of employment/occupation, income, household size, health insurance (yes/no), tobacco use, and alcohol use.

2.2.2. Neighborhood level measures

Neighborhood was defined as census tract of residence.

Food environment variables: we defined and examined the impact of the neighborhood food environment as the density of specific types of food establishment (e.g., outlets per capita) using a continuous scale. Types of food establishments were derived from 2011 InfoUSA data. From listed grocery stores, two subsets were characterized – large grocery stores that typically sell fresh foods and convenience stores including local and national chains (e.g., "Seven-Eleven"). Fast food establishments were identified from listed fast food chain restaurants. In addition, any restaurant or convenience store whose name included fried chicken, sandwich, fries, burgers, hot dogs, shakes, pizza, drive through, and express was added to the unhealthy food outlet category. Further, all outlets listed as bars in the InfoUSA data base were identified. Finally, the outlet densities (i.e., outlets per census population) were characterized into three indices of food and beverage outlets: (1) healthy outlet density which included counts of large grocery stores as the numerator, (2) unhealthy outlet density which included the count of fast food outlets and convenience stores (Rundle et al., 2009) as the numerator, and (3) bar density which included the count of all outlets listed as bars as the numerator.

Physical activity environment variables: as with food establishment data, commercially available data from InfoUSA were used to characterize availability of physical activity conducive facilities using the SIC codes which correspond to those used by the US Census (Gordon-Larsen et al., 2006; Nelson et al., 2006). A list of SIC codes representing physical activity related facilities were compiled and used to identify and enumerate those facilities, including parks. The variable was characterized as the density of physical activity facilities.

Walkability variables: walkability was defined as the degree of street connectivity in a census tract (i.e., neighborhood street networks that are continuous, integrated, and maximize linkages between starting points and destinations, providing multiple route options) (Greenwald and Boarnet, 2001; Saelens et al., 2003). The indices of street connectivity calculated for the analysis are described in Table 1. They include intersection density, street density, and connected node ratio. Intersection density refers to the density of intersections in a neighborhood (i.e., more intersections per unit area, more connectivity). Street density refers to the degree to which a neighborhood has a high concentration of streets (i.e., more street miles per unit area, more connectivity). Finally, connected node ratio refers to the degree to which intersections in a neighborhood are of the types that increase connectivity (i.e., more four way intersections yield more connectivity).

Population density was defined as total population per unit area (minus commercial, industrial, and park land) based on US Census estimates, with a higher percentage representing a higher population density.

Crime variable: the degree to which a neighborhood was exposed to crime was obtained from the ESRI. The ESRI data provides an index of violent crime at the census tract level. The index is generated by modeling city or county crime statistics to infer rates at the census tract level.

Economic deprivation variables: economic deprivation was measured using two different variables: (1) income-to-poverty ratio and (2) concentrated disadvantage index. The income-to-poverty ratio is assessed at the individual level in the NHANES. It represents "the ratio of family or unrelated individual income to their appropriate poverty threshold. Ratios below 1.00 indicate that the income for the respective family or unrelated individual is below the official definition of poverty, while a ratio of 1.00 or greater indicates income above the poverty level" (U.S. Census Bureau, 2004). The ratio was categorized into tertiles as low, medium, and high. An index of concentrated disadvantage (Sampson and Morenoff, 2004) was generated at the census tract level. Concentrated disadvantage is derived from six census measures including the percent of families or households below the poverty line, percent of families receiving public assistance, percent of unemployed individuals in the civilian labor force, percent of population that is black, percent of population less than 18 years of age, and percent of families with children that have a female as the head of the household.

2.3. The mediation analysis

In order to explore the factors that can explain the racial disparity in obesity we adapted the mediation analysis method proposed by Yu et al. (2014). The method was implemented using the mma package in the statistics software R (Yu and Li, 2017). Note that the purpose for this study is not to identify a causal relationship, but to explore the racial disparity. That is, the purpose was to identify variables that can explain the racial disparity, but not assume that those variables cause the racial disparity. However, the inferences for the two concepts are statistically identical (MacKinnon et al., 2000). In this paper, mediators are referred to as the related factors that can be used to explain the observed relationship between an

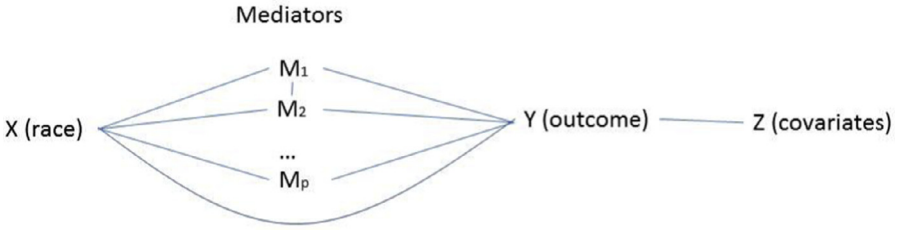

Fig. 1. Graphical relationship among variables.

exposure variable and an outcome. The exposure variable is race (black vs. white) and the outcome is the binary variable obesity. The relationship among variables is pictured in Fig. 1, where mediators (denoted as M) are those variables highly related to both the exposure and outcome variable, and Z are the covariates that are related to the outcome but not the exposure variable. In Yu et al.'s paper, the total effect from race X on outcome Y is defined as $TE_Z = E(Y|Z, X = 1) - E(Y|Z, X = 0)$, where Z are other covariates and $X = 1$ for blacks and 0 for whites. The total effect describes the total racial disparity in the outcome. The direct effect not from a certain mediator M_j is defined as

$$DE_{\setminus Mj|Z} = E_{m_j}\left\{E_{M_{-j}|X=1}\left[E\left(Y|Z, M_j = m_j, M_{-j}, X = 1\right)\right] - E_{M_{-j}|X=0}\left[E\left(Y|Z, M_j = m_j, M_{-j}, X = 0\right)\right]\right\},$$

where M_{-j} indicate all mediators excluding the jth mediator, M_j. The definition implies that DE measures the average difference in the outcome where M_{-j} follows the conditional distribution on $X = 0$ or 1 while M_j are fixed at its marginal distribution that does not change with X. Intuitively, by manipulating the values of M_j, we de-correlate the association of X with M_j, and therefore remove the indirect effect from X to M_j to the outcome. The indirect effect of X on Y through M_j is thus defined as the difference between $TE_{|Z}$ and $DE_{\setminus Mj|Z}$. This novel mediation analysis methodology was adopted for four reasons. First, the method is general so that the related factors in study can be measured in different scales: continuous, binary or multi-categorical. Second, multiple factors of different types are allowed in the pathway analysis simultaneously. Indirect effect transmitted by an individual factor or by a subset of factors can be differentiated from the total effect, which enables the comparison of the importance of the individual or joint factors. With the knowledge of the indirect effect carried by each mediator/confounder in the racial disparity in obesity, a policymaker is potentially able to focus limited resources on the most important factors to reduce the racial disparity efficiently. Third, the mediation study allows correlations among factors. In addition to differentiate indirect effect from each risk factor, a group of variables can be considered jointly. Therefore, the indirect effect through neighborhood walkability, which was measured by three related variables (see Table 1), can be measured. Fourth, the concepts of mediation analysis can be applied in general predictive models, so that in addition to the generalized linear models (GLM), more complex models can be formulated to account for complicated relationships among variables. Finally, Yu et al. (2014) provided two approaches (the Delta method and bootstrap method) to estimate the variances of the estimates of mediation effects with parametric or nonparametric models, which makes the inferences on mediation effects possible.

The outcome considered in this paper is obesity (binary). All other factors described in the *measures* section were considered as candidate mediators. Only those variables that satisfied the two conditions of being: (1) significantly related to race, and (2) significantly related to the outcome variable after adjusting for other factors, were included in the final model as potential mediators. Those variables that satisfied condition (2) but not (1) were included in the final model as other covariates. To select the variables, we choose a significance level of 0.1 for inclusion. The R function "data.org" in mma package was used to identify all potential mediators. The function returns a reorganized data set, which includes only the important mediators, predictors, and covariates. In the study, one goal was to explore the effects of walkability. However, walkability was measured by three variables listed in Table 1; these variables were highly correlated. The goal was to determine the joint effect from these measures instead of any individual effect. Therefore, these variables were forced to enter the model and their joint effects were estimated.

We used two predictive models to model the association between obesity and all other predictors. One is the generalized linear model (logistic regression for obesity) and the other is the MART (Bernoulli distribution for obesity). Compared with the classical parametric regression methods, MART has the following advantages: (1) MART is able to capture the nonlinear relationships between the dependent and independent variables with no need for specifying the basis functions. (2) Because of the hierarchical splitting scheme in regression trees, MART is able to capture high-order interaction effects and the hierarchical data structure at both individual and census tract levels. (3) Unlike many automated learning procedures, which lack interpretability and operate as a "black box", MART provides tools to interpret the nature and magnitudes of covariate relations with the outcome (for example, relative variable importance and partial dependence plots) (Friedman, 2001; Yu et al., 2009). (4) MART can handle mixed-type predictors (i.e. quantitative and qualitative covariates) and missing values in covariates. (5) MART has shown a superior exploration and prediction performance in epidemiology research (Friedman and Meulman, 2003;

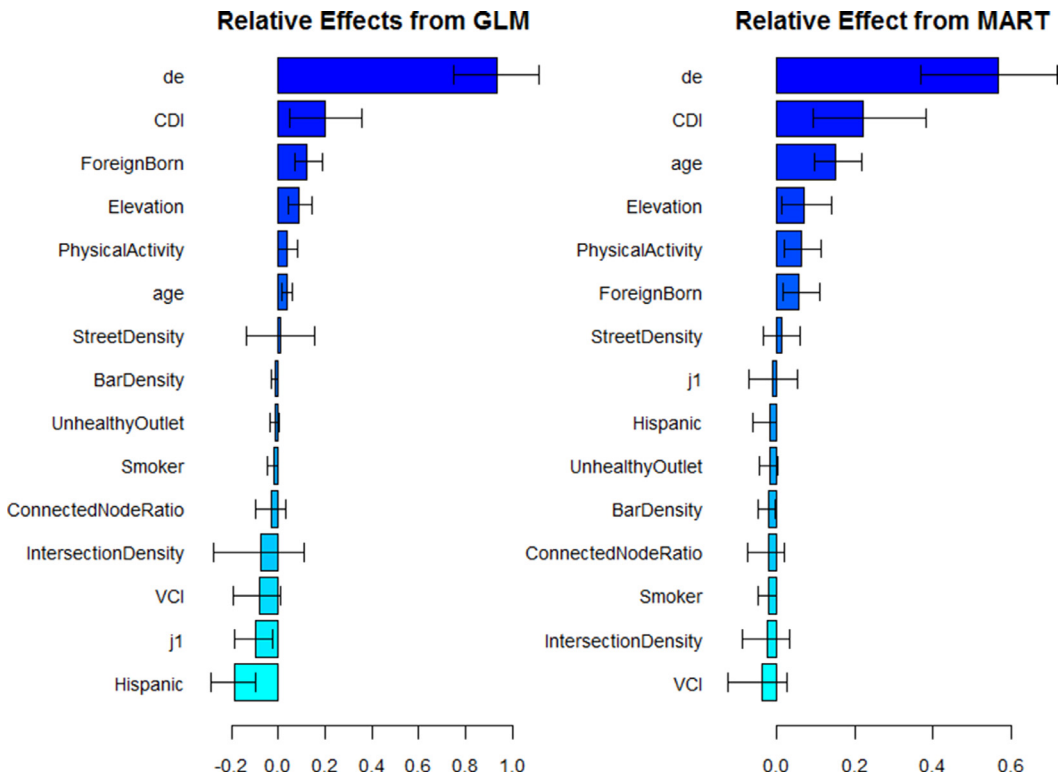

Fig. 2. Estimated relative effects by GLM (left) and MART (right).

Yu et al., 2009). The R functions "boot.med" and "mma" in the *mma* package (Yu and Li, 2017) were used to make inferences on the mediation effects. The plot functions included in the *mma* package were then adapted to explore the complicated variable associations.

3. Results

The descriptive analysis, presented in Table 2, reveals a number of differences between the blacks and whites in the continuous NHANES sample used for the analyses. With regard to individual level variables blacks are more likely to be younger, US born, smokers, and have low poverty to income ratios, higher energy intake, and higher sugar sweetened beverage consumption. With regard to the census tract level variables, blacks are more likely to live in census tracts characterized by lower elevations, lower density of unhealthy food outlets, higher violent crime indices, higher concentrated disadvantage, lower bar density and lower grocery store density and higher measures of walkability. The classification of Hispanics as white, which represents 29.0% of the whites sampled, undoubtedly accounts for the greater likelihood of whites being classified as foreign born. The census tract level variables implicated in explaining the disparity in obesity rates between blacks and whites would appear to be residence in neighborhoods characterized by high levels of concentrated disadvantage, a higher level of crime, low elevation, and low grocery store density. Conversely, the tract variables which go against any disparity in obesity rates between blacks and whites would be the residence of blacks in neighborhoods characterized by higher walkability and the lower unhealthy food outlet density than those of whites.

We next consider the results of the mediation analyses. Using the rules described in Section 2.3, variables were included as covariates but not potential mediators if they were associated with the outcome but not distributed differently by race for each outcome. The three measures of walkability were forced to enter the predictive models as potential joint mediators. Table 2 lists the summary statistics of all variables that were selected in the final models classified by race. Analysis of variance (ANOVA) was used to check the association between race and each continuous variable, and χ^2 test were used to check that between race and each categorical variable. As a result, the variables elevation, population density, physical activity, violent crime index, concentrated disadvantage index, gender, race, poverty category, foreign born status, smoking status, total calories, age, Hispanic, unhealthy food outlet density, and the joint walkability measurement were adopted in the final model for explaining obesity.

Table 3 lists the inferences on the mediation effects for only the variables with a significant indirect effect with either the GLM or MART model. Both logistic regression and MART with a logit link were used as the predictive models to permit comparison of the two methods. The relative effect (RE) is defined as the ratio of the corresponding (in)direct effect over the total effect. A positive relative effect means that a portion of the racial disparity is

Table 2
Summary statistics for variables used in the final models by race from the NHANES ($N = 5240$).

Continuous variables	Black	White	*P*-value
	Means (SD)	Means (SD)	ANOVA
Individual level variables			
BMI[1], kg/m[2]	30.10(6.8)	28.06(5.8)	<.0001
Age, years	49.76 (16.33)	53.02 (18.43)	<.0001
Physical activity[2]	0.67 (1.54)	0.74 (1.52)	0.1447
Census tract level variables			
Elevation, m	150.91 (231.75)	333.34 (412.54)	<.0001
Population density, count/mile[2]	8029.88 (11,483)	4381.39 (8128)	<.0001
Unhealthy outlet density, count/1000 people	0.98 (1.08)	1.10 (1.03)	0.0005
Violent crime index	226.19 (119.71)	105.14 (115.83)	<.0001
Concentrated disadvantage index	0.98 (1.26)	−0.16 (0.78)	<.0001
Bar density, count/1000 people	0.15 (0.36)	0.17 (0.31)	0.0837
Grocery store density, count/1000 people	0.49 (0.53)	0.55 (0.51)	0.0028
Walkability measures			
Street density, mile/mile[2]	12.43 (7.18)	9.28 (6.84)	<.0001
Intersection density, count/mile[2]	97.95 (77.34)	62.92 (65.10)	<.0001
Connected node ratio, count/count	0.80 (0.14)	0.75 (0.13)	<.0001
Categorical variables	Proportion (%)	Proportion (%)	χ^2 test
Individual level variables			
Obese[1]	44.32	30.23	<.0001
Male[2]	50.29	51.80	0.39
Foreign born	7.63	23.97	<.0001
Current smoker	22.60	19.30	0.02
Hispanic	0	29.02	<.0001
Family poverty income ratio			
<1x poverty level	16.54	14.44	0.02
1x–3x poverty level	43.44	40.97	
>3x poverty level			
Total energy intake tertile, kcal			
Lower	32.97	33.43	0.007
Medium	29.94	34.16	
Higher			
Sugar and sweet beverages intake tertile, kcal			
Lower	28.57	22.07	<.0001
Medium	27.69	22.29	
Higher			

[1] Outcome variable.
[2] Variable used as covariate but not as potential mediator.

Table 3
Racial disparity in obesity risk explained by other factors.

Factors with significant indirect effect	Logistic regression		MART	
	Effects (sd)	RE (%)	Effects (sd)	RE (%)
Elevation	0.06 (0.02)	8.9	0.04 (0.02)	7.0
Concentrated disadvantage index	0.13 (0.05)	20	0.12 (0.03)	22.1
Physical activity	0.02 (0.01)	3.8	0.03 (0.01)	6.3
Age	0.02 (0.01)	3.5	0.08 (0.01)	5.6
Nativity status = foreign born	0.08 (0.02)	12.1	0.03 (0.01)	5.56
Hispanic ethnicity	−0.12 (0.03)	−18.5	0.03 (0.01)	−1.6
Smoking status = current smoker	−0.01 (0.01)	−2.2	−0.01 (0.01)	−2.1
Unhealthy outlet density	−0.01 (0.01)	−1.6	−0.01 (0.01)	−1.6
Bar density	−0.01 (0.02)	−1.3	−0.01 (0.01)	−2.0
Joint effect: street connectivity	−0.06 (0.02)	−9.6	0.01 (0.02)	−0.9
Unexplained racial disparity	0.60 (0.08)	93.2	0.31 (0.08)	56.67
Total racial disparity	0.64 (0.08)	100	0.54 (0.08)	100

accounted for by the factor. A negative relative effect means that instead of explaining the racial disparity, accounting for the corresponding factor is associated with an enlargement of racial disparity in the outcome.

Fig. 2 compares the importance of all potential mediators in explaining the racial disparity in obesity in terms of their relative effects. Results from logistic regression and MART are compared side by side. With regard to explaining racial disparity in obesity risk, relative effects for several individual and census tract level variables were consistent between the two models. At the individual level the relative effects for differences between

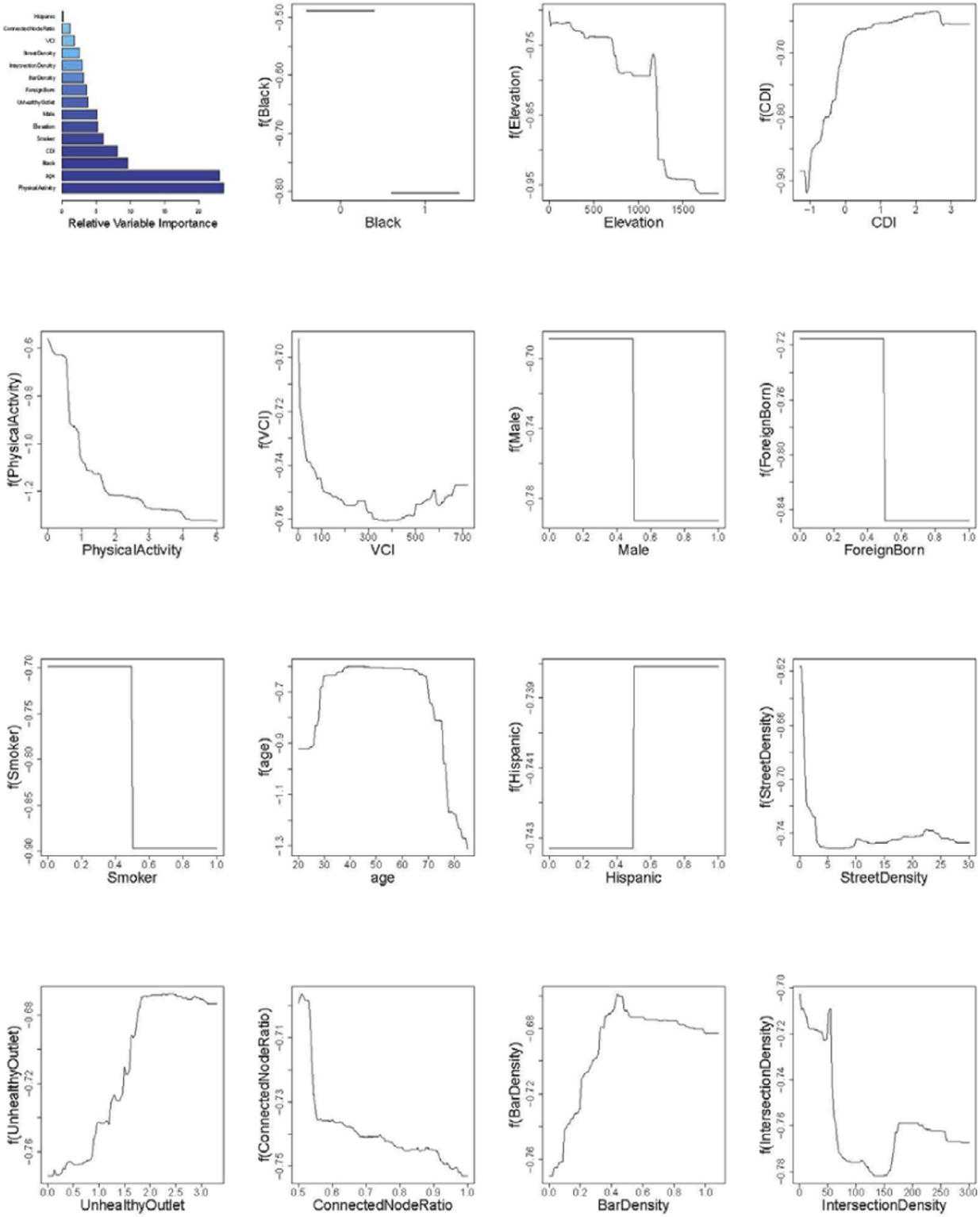

Fig. 3. Relative importance and partial dependence plots of variables from the Multivariate Additive Regression Trees (MART) obesity model.

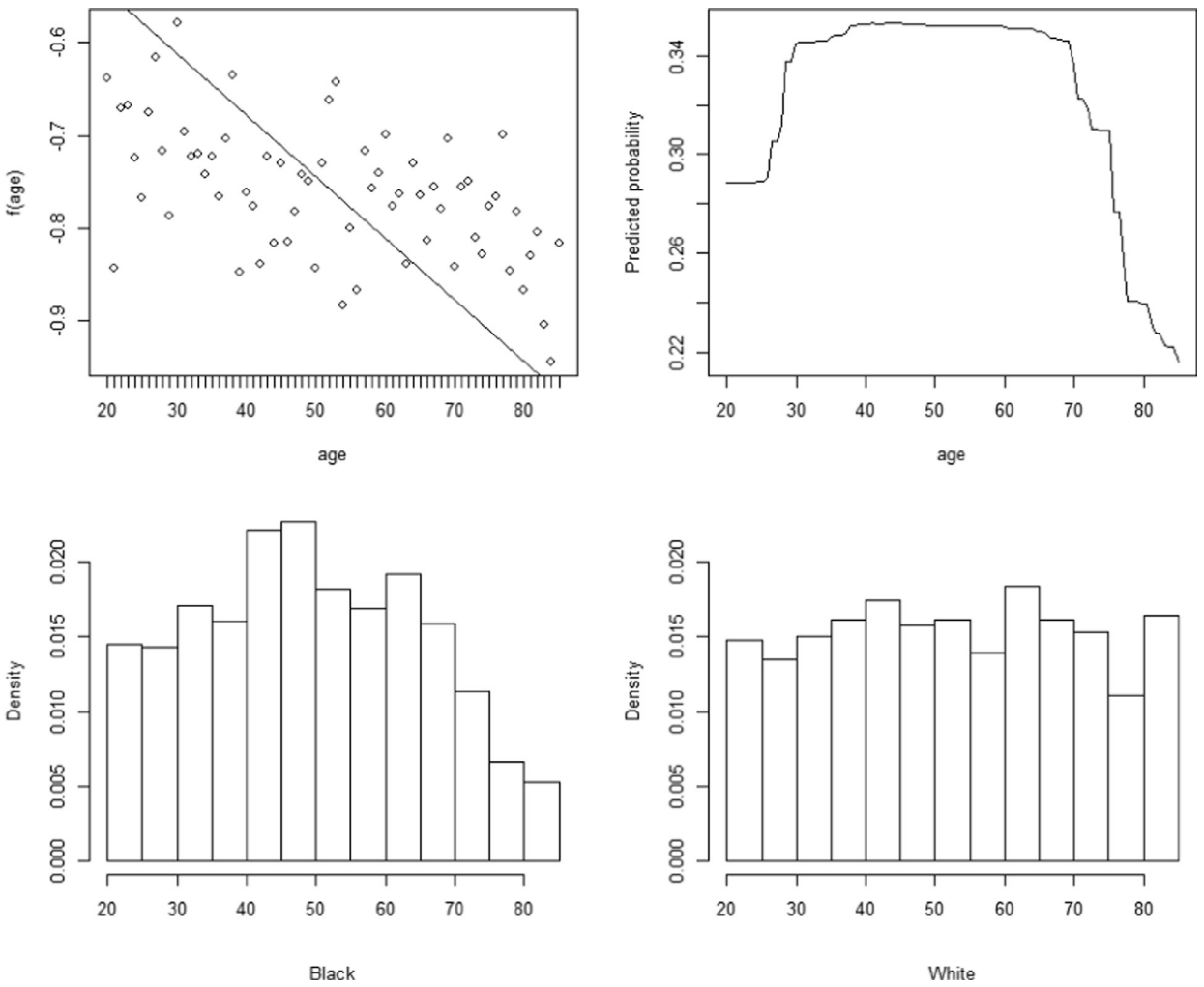

Fig. 4. The upper panel shows the fitted relationship between age and the probability of obesity by logistic regression (left) and MART (right). The lower panel shows the age distribution within black (left) and white (right) population.

blacks and whites in age distribution and nativity status explained some of the disparity in obesity risk in both models. At the census tract level, the relative effects of differences between blacks and whites in terms of residing in census tracts characterized by concentrated disadvantage, bar density, and elevation explained some of the disparity in both models. Inconsistencies between the two models were noted for other variables. A negative relative effect for physical activity was observed in the MART model but no significant relative effect was observed in the logistic regression. The joint effect of the walkability measures ($j1$) in terms of its relative effect showed a difference between the two models. The logistic regression model accounting for walkability evidenced a negative relative effect (i.e., the disparity was exacerbated due to inclusion), while the MART accounting for street connectivity evidenced no significant relative effect accounting for other variables. It should be noted that due to the large negative effects for Hispanic ethnicity and walkability, only 7% of the net disparity in obesity risk was explained in the logistic regression model while the MART model explained about 43% of the disparity.

Fig. 3 shows the relative importance and partial dependence plots of some important variables from the MART models for obesity. It is interesting to note that the dependence plots for the relationship between two of the street connectivity measures (i.e., intersection density and street density) demonstrate a non-linear relationship. Specifically, at the higher levels of intersection density and street density, the relationship with obesity is no longer negative but becomes positive.

4. Conclusion and discussion

Although there were many similarities between GLM and MART models for describing the factors mediating the disparity between black race and both obesity and BMI, there were notable differences. Overall the MART models appeared to perform better in explaining the racial disparity with 43.4% explained in the obesity model. The

difficulty for the GLM came with variables like street connectivity, and Hispanic ethnicity which suppressed the ability of the models to explain the disparity. The capability of the MART model to address the possibility of nonlinearity for variables like street connectivity contributed to its increased explanatory power.

The results from MART and the GLM mostly agreed in terms of the variables selected as important mediators and the directions of their indirect effects. However, there were also many differences in terms of the magnitude of the relative effects from each mediator for the two kinds of models.

The variables with similar effects for both modeling approaches in explaining the disparity in obesity risk were elevation of residence, residing in a neighborhood characterized by concentrated disadvantage, and foreign nativity status. As shown in Fig. 3, compared with whites, blacks are more likely to reside at higher elevations, reside in neighborhoods characterized by concentrated disadvantage, and be US born. All these factors have been associated with increased risk of obesity. The relative effect from age is more complicated. The variable age explained 15% of the racial disparity in obesity in the MART model, but only 3.5% by logistic regression. Fig. 4 shows the fitted relationship between the obesity risk and age by logistic regression and MART separately. It also presents the age distributions among blacks and whites separately. The MART dependence plot shows that age and the probability of obesity did not have a linear relationship. The probability of obesity increased with age until around 40, at which point the proportion of obesity reached its maximum and remained largely steady until around 70 years of age, where it began to decrease. Among the subjects in the study, 64% of whites were between the ages of 30 and 70, compared to 74% for blacks. Since MART detected the nonlinear relationship between age and obesity risk, it explained more of the racial disparity in obesity through age compared with logistic regression.

The relative importance and partial dependence plots in Fig. 3 provide insights into the nature of the relationships between the various variables and obesity risk in MART. The first plot of Fig. 3 shows the relative importance of different variables in explaining the variances in obesity. The MART model indicates that after adjusting for other variables, race was still an important predictor in explaining obesity risk. The rest are the partial dependence plots showing the relationship between obesity and each of the important variables. We found that, on average, the probability of obesity decreased as the elevation increased. Also, the average elevation was higher for whites (333) than for blacks (413). In view of this, the finding that elevation explained 7% of the racial disparity in obesity reported in Table 3 is unsurprising. Considering the concentrated disadvantage index (CDI), blacks had a higher average score than whites (0.98 vs. −0.16), while the probability of obesity increased with CDI. CDI thus explained 22.1% of the racial disparity. We also found that foreign born residents were less likely to be obese. However, whites were more likely to be foreign born than blacks (23.8% vs. 7.6%). Being foreign born thereby explained 5.6% of the obesity. The explanation of the joint effect of walkability is complicated since it involved three variables (Street Density, Intersection Density and Connected Node Ratio) and the relationships between those variables and the probability of obesity were complex as described in the dependence plot. Average walkability was higher for blacks and on average, higher walkability is related with less obesity. Therefore, the relative effect is negative by logistic regression, implying that accounting for walkability would only enlarge the racial disparity. However, the relative effect of walkability is not significant by MART. This may be due to the nonlinearity seen in the partial dependence plots for street and intersection density, the risk of obesity declined as densities increased but then rose again at the higher densities. This could reflect a greater black presence in environments less conductive to walking despite relatively high connectivity.

In addition to the mediators that can be used to explain increases in apparent racial disparities, others acted as suppressors that reduced apparent disparity. For example, smoking was related to less obesity while a higher proportion of blacks smoked than whites (22.6% vs. 19.3%). This factor actually reduced the apparent racial disparity in obesity, accounting for −2.1% of it. In the same vein, an increased unhealthy outlet density was related to an increased proportion of obesity while blacks had a smaller index (0.98) compared with whites (1.1). The relative effect of the unhealthy population index was thus −1.6%. A similar association was found for the factor bar density.

There are some limitations to the analysis that need to be mentioned in considering these results. First is the likelihood that there are omitted variables that might alter the results and increase the explanatory power of the models. Second, as previously noted, the use of cross-sectional survey data allows conclusions only about associations, not causality. Finally, census tract level variables such as those included in the analysis are ideally addressed in a multilevel model. This technique may have yielded differing results, but is not currently available for MART-based mediation analyses. For the future, effort should be directed to the development of methods for mediation analysis using multilevel models.

Authors' contributions

All authors contributed to the development of study concepts. QY, RS, CN, LZ, LC, CP and NS participated in creating neighborhood level variables from census data and merged them with the NHANES dataset. QY, CN and RS conducted the statistical analyses and drafted first version of the manuscript. All authors reviewed and edited on the manuscript.

All authors have approved the final paper.

Funding

This work was supported by the National Cancer Institute (NCI) (grant number R01 CA157565); and the Seed Fund, Louisiana State University Health Sciences Center. Both funding bodies do not influence the design of the study, collection, analysis, and interpretation of data, and the manuscript writing.

Conflicts of interest

None.

References

Albert JM. Mediation analysis via potential outcomes models. Stat Med 2008;27:1282–304.

Ball K, et al. Street connectivity and obesity in Glasgow, Scotland: impact of age, sex and socioeconomic position. Health Place 2012;18:1307–13.

Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol 1986;51:1173–82.

Berrigan D, Troiano RP. The association between urban form and physical activity in US adults. Am J Prev Med 2002;23:74–9.

Ding D, Gebel K. Built environment, physical activity, and obesity: what have we learned from reviewing the literature? Health Place 2012;18:100–5.

Eriksson U, et al. Walkability parameters, active transportation and objective physical activity: moderating and mediating effects of motor vehicle ownership in a cross-sectional study. Int J Behav Nutr Phys Act 2012;9:1–10. Available at http://dx.doi.org/10.1186/1479-5868-9-123.

Frank LD, et al. A hierarchy of sociodemographic and environmental correlates of walking and obesity. Prev Med 2008;47:172–8.

Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001:1189–232.

Friedman JH, Meulman JJ. Multiple additive regression trees with application in epidemiology. Stat Med 2003;22:1365–81.

Gebel K, Bauman AE, Petticrew M. The physical environment and physical activity: a critical appraisal of review articles. Am J Prev Med 2007;32:361–9 e3.

Gordon-Larsen P, et al. Inequality in the built environment underlies key health disparities in physical activity and obesity. Pediatrics 2006;117:417–24.

Grasser G, et al. Objectively measured walkability and active transport and weight-related outcomes in adults: a systematic review. Int J Public Health 2013;58:615–25.

Greenwald M, Boarnet M. Built environment as determinant of walking behavior: analyzing nonwork pedestrian travel in Portland, Oregon. Transp Res Rec J Transp Res Board 2001:33–41.

Harris K. Access to confidential data for statistical analysis. Bethesda, MD, U.S: Department Of Health And Human Services, Centers For Disease Control And Prevention, National Center for Health Statistics, NCHS, Research Data Center; 2006 NCHS Meeting.

Have TR, et al. Causal mediation analyses with rank preserving models. Biometrics 2007;63:926–34. Available at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17825022.

Heinrich KM, et al. How does the built environment relate to body mass index and obesity prevalence among public housing residents? Am J Health Promot 2008;22:187–94.

Hill JO, et al. Obesity and the environment: where do we go from here? Science 2003;299:853–5.

Khan LK, et al. Recommended community strategies and measurements to prevent obesity in the United States. MMWR Recomm Rep 2009;58:1–26.

King K. Neighborhood walkable urban form and C-reactive protein. Prev Med 2013;57:850–4.

Li F, et al. Multilevel modelling of built environment characteristics related to neighbourhood walking activity in older adults. J Epidemiol Community Health 2005;59:558–64.

Li F, et al. Built environment, adiposity, and physical activity in adults aged 50–75. Am J Prev Med 2008;35:38–46.

MacKinnon DP, Krull JL, Lockwood CM. Equivalence of the mediation, confounding and suppression effect. Prev Sci 2000;1:173–81.

MacKinnon DP, Warsi G, Dwyer JH. A simulation study of mediated effect measures. Multivar Behav Res 1995;30:41–62.

McDonald KN, Oakes JM, Forsyth A. Effect of street connectivity and density on adult BMI: results from the Twin Cities walking study. J Epidemiol Commun Health 2012;66:636–40.

Nelson MC, et al. Built and social environments: associations with adolescent overweight and activity. Am J Prev Med 2006;31:109–17.

Oakes JM, Forsyth A, Schmitz KH. The effects of neighborhood density and street connectivity on walking behavior: the Twin Cities walking study. Epidemiol Perspect Innov 2007;4:1.

Owen N, et al. Neighborhood walkability and the walking behavior of Australian adults. Am J Prev Med 2007;33:387–95.

Papas MA, et al. The built environment and obesity. Epidemiol Rev 2007;29:129–43.

Pearce JR, Maddison R. Do enhancements to the urban built environment improve physical activity levels among socially disadvantaged populations? Int J Equity Health 2011;10:1.

Pearl J. Direct and indirect effects. Proceedings of the seventeenth conference on uncertainty and artificial intelligence. CA: Morgan Kaufmann, 2001.

Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology 1992;3:143–55. Available at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=1576220.

Roux AVD. Residential environments and cardiovascular risk. J Urban Health 2003;80:569–89.

Rundle A, et al. Neighborhood food environment and walkability predict obesity in New York City. Environ Health Perspect 2009;117:442.

Saelens BE, Handy SL. Built environment correlates of walking: a review. Med Sci Sports Exerc 2008;40 S550.

Saelens BE, Sallis JF, Frank LD. Environmental correlates of walking and cycling: findings from the transportation, urban design, and planning literatures. Ann Behav Med 2003;25:80–91.

Sampson RJ, Morenoff JD. Spatial (dis)advantage and homicide in Chicago neighborhoods. In: Goodchild M, Janelle D, editors. Spat Integr Soc Sci. New York: Oxford University Press; 2004. p. 145–70.

Wang F, Wen M, Xu Y. Population-adjusted street connectivity, urbanicity and risk of obesity in the US. Appl Geogr 2013;41:1–14.

Wen M, Kowaleski-Jones L. The built environment and risk of obesity in the United States: racial–ethnic disparities. Health Place 2012;18:1314–22.

Witten K. Neighborhood built environment and transport and leisure physical activity: findings using objective exposure and outcome measures in New Zealand. Environ Health Perspect 2012;120:971.

Yu Q, Li B. mma: an R package for mediation analysis with multiple mediators. J Open Res Softw 2017, in press.

Yu Q, Fan Y, Wu X. general multiple mediation analysis with an application to explore racial disparity in breast cancer survival. J Biomet Biostat 2014;5:189.

Yu Q, Li B, Scribner RA. Hierarchical additive modeling of nonlinear association with spatial correlations–an application to relate alcohol outlet density and neighborhood assault rates. Stat Med 2009;28:1896–912. Available at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=19402025.