AJE Advances: Research in Epidemiology, 2025, **1(3)**, uuaf011
https://doi.org/10.1093/ajeadv/uuaf011
Advance access publication: September 24, 2025
Original Contribution

Neighborhood environmental and contextual factors improve prediction of childhood body mass index: an application of novel graph neural networks

Keyu Li[1], Charles Wood[2], Liz Nichols[3], Zachary D. Calhoun[4], Nrupen A. Bhavsar[*,†,3,5] and David Carlson[†,1,4,5]

[1]Department of Electrical and Computer Engineering, Duke University Pratt School of Engineering, Durham, NC 27705, United States
[2]Department of Pediatrics, Duke University School of Medicine, Durham, NC 27705, United States
[3]Department of Surgery, Duke University School of Medicine, Durham, NC 27705, United States
[4]Department of Civil and Environmental Engineering, Duke University Pratt School of Engineering, Durham, NC 27705, United States
[5]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27705, United States
[*]Corresponding Author. Nrupen A. Bhavsar, Division of Surgical Sciences, Department of Surgery, Duke University School of Medicine, 710 W. Main Street, Durham, NC 27701, United States (nrupen.bhavsar@duke.edu)
[†]Nrupen A. Bhavsar and David Carlson are senior authors and have contributed equally

Abstract

Childhood obesity is a major risk factor for adult cardiovascular disease. Current obesity-prediction models were not developed in diverse populations and do not include heterogeneous social, environmental, and climate factors that may impact body mass index across the full pediatric spectrum. Additionally, they consider only the immediate neighborhood within which a child lives, ignoring contextual factors from expanded (ie, distal) neighborhoods. This study uses expanded neighborhoods' social, environmental, and climate data to improve individual-level body mass index prediction–from underweight through obesity–using a novel machine learning approach. We obtained demographic and clinical data from the electronic health records of the Duke University Health System, identifying 12,226 children aged 6-18 years in Durham County, North Carolina, with body mass index data from 2014 to 2022. Participants' data were linked to socioeconomic and environmental information at the census block group level. We captured expanded neighborhood effects with a graph neural network and combined this information with individual-level factors to predict body mass index. Our model predicted body mass index more accurately than simpler models for children aged 6–11 ($R^2 = 0.234$, mean absolute error = 3.352, root mean square error = 4.370) and 12–18 ($R^2 = 0.147$, mean absolute error = 4.980, root mean square error = 6.343) using all features. Key predictive factors identified included rent burden, poverty rate, and tree coverage. This research highlights the value of including broader socioeconomic and environmental factors in body mass index prediction, offering insights that could guide targeted, neighborhood-level interventions.

Key words: electronic health records; geospatial modeling; child exposure/health; machine learning.

Introduction

Obesity in children is strongly associated with cardiometabolic risk factors such as hypertension, insulin resistance, and dyslipidemia and increases the risk for diabetes, cardiovascular disease, and premature mortality.[1-4] Approximately 80% of adolescents with obesity will continue to have obesity into adulthood, resulting in substantially higher risk for cardiovascular disease as adults.[5] The factors that predispose children to obesity include household and individual factors, and social and environmental determinants of health (SDOH).[6,7] Identifying the specific social and environmental factors that predict childhood obesity may improve prevention by helping prioritize policy interventions.

Existing models to predict childhood obesity incorporate demographic, clinical, and parent-level factors.[8] However, these models have largely been developed and validated in cohort studies of homogenous European populations and are not necessarily generalizable to the US population, which is more racially and socioeconomically diverse**.** In the US, the prevalence of childhood obesity is significantly higher among non-Hispanic Black and Hispanic children compared to non-Hispanic white children.[4] Excluding diverse populations when developing childhood obesity prediction models may lead to poor predictions in those groups and incorrect clinical decisions.[9] Also, current models that predict individual BMI primarily include clinical and demographic variables, with limited or no neighborhood level environmental or contextual factors. Even when such factors are included, they are often restricted to a person's immediate surroundings and fail to capture influences from broader communities.

When considering how SDOH can improve prediction characteristics, we are mindful of the limitations of current measures of spatial autocorrelation. Traditional spatial autocorrelation approaches rely on parametric, linear combinations of predictors while SDOH predictors may be non-linear and correlated.[10] These approaches also require that data are observed.[10] Novel

Received: May 2, 2025. **Accepted:** September 3, 2025

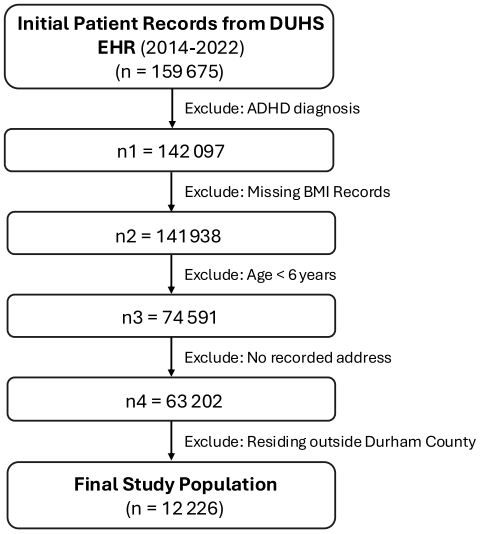
machine learning approaches can be designed to better account for autocorrelation between expanded neighborhoods.

This study addresses these limitations by linking electronic health record (EHR) data from a health system that serves a racially, ethnically, and socioeconomically diverse patient population with social, environmental, and climate data, focusing on incorporating the geographic distribution of expanded neighborhood environments. We propose a custom graph neural network (GNN) approach that captures autocorrelation between neighborhood-level SDOH variables and combines it with individual-level clinical data. This is used to develop a novel framework to predict body mass index (BMI) in children, which is used to risk stratify children based on categories (ie, overweight, obesity, and severe obesity). The resulting model has the potential to more accurately predict childhood obesity and identify opportunities for prevention efforts.

Methods

Data sources and study population

The study population and clinical data were obtained from the EHR of the Duke University Health System (DUHS). DUHS is the primary provider of healthcare in Durham County, NC. Durham County includes racially, ethnically, and socioeconomically diverse urban, suburban, and rural areas and nearly 90% of children in Durham receive care within DUHS.[11] In our study, children were eligible for the study if they had at least two encounters with body mass index (BMI) recorded between 9 to 36 months apart. We only used the first BMI measurement in our study. The resulting source population for this study included 159,675 children under 18 years of age residing within North Carolina from 2014-2022. Children were excluded if they were less than 6 years of age to focus on school-aged children and adolescents, if they lacked a recorded address, were missing BMI value, lived outside of Durham County, or were diagnosed with attention deficit hyperactivity disorder (due to stimulant effects on BMI). As a result, 147,449 children were excluded from the study. Patient addresses were geocoded and assigned FIPS codes at the block group level. This study was approved by the Duke University School of Medicine Institutional Review Board. A flowchart outlining the inclusion and exclusion criteria used to define the final study population is presented in Figure 1.



BMI = Body Mass Index; EHR = Electronic Health Records

Figure 1. Flowchart of exclusion criteria used to define the final study population.

Socioeconomic, built environment, and climate predictors

Block group–level socioeconomic, built environment, and climate factors from 2014-2022 were included in the model based on their known associations with childhood obesity.[12-17] Socioeconomic and built environment data were obtained from the Social, Environmental, and Equity Drivers (SEED) Health Atlas,[12] Durham Neighborhood Compass,[18] DataAxle,[19] United States Environmental Protection Agency,[20] North Carolina State Climate Office,[21] US Geological Survey,[22] Google Earth Engine[23] and prior research.[24] These organizations provide valuable data on social and environmental factors, and the built environment, contributing to the comprehensive scope of SDOH data on the SEED Health Atlas, a web based platform that democratizes data. Socioeconomic variables included 2014-2022 Childhood Opportunity Index (COI) (block groups were assigned the COI of their containing census tract), median household income, unemployment rate, percent of residents without a vehicle, percentage of households with high rent burden, percent of individuals living at or below poverty level, and percentage of households with more than one person per room. Built environment variables included 2014-2022 neighborhood tree coverage, National Walkability Index, LandSat Normalized Difference Vegetation Index (NDVI), and data from DataAxle on the number of liquor stores, grocery stores, convenience stores, religious institutions, and fast-food restaurants within a block group.

To estimate neighborhood-level variations in climate, we used block-group level estimates of summertime evening air temperature to serve as an indicator of the urban heat island effect. To produce these estimates for 2014-2022, we adapted a land-use and vegetation-based model following an established method.[25] This model is based on data collected from the NOAA (Office of Education, Climate Program Office, National Integrated Heat Health Information System (NIHHIS)) heat island mapping campaign in July 2021, in which evening temperature was observed at high spatial resolution.[21] We fit the model using 2021 National Land Cover Dataset (NLCD) and NDVI data, then used the model to estimate neighborhood evening temperature based on NLCD and NDVI data from the other years (2014-2020, 2022). Since this method produced rasterized estimates for each year at 30-meter resolution, we then averaged the estimates over census block group boundaries for use in the GNN. We provide several example visualizations of temperature over time in supplementary material (Figure S1).

Demographic and clinical data

Demographic predictors included patient age and sex. Clinical data was used to quantify BMI (the outcome), which was defined using weight and height (ie, kg/m^2).[26]

Geographic information

We collected TIGER/Line shapefiles including geographic and cartographic information of Durham County in 2014-2022 from United States Census Bureau.[27] Patient level residential address and associated block group level Federal Information Processing Standards (FIPS) codes at the block group level were obtained from DUHS EHR.

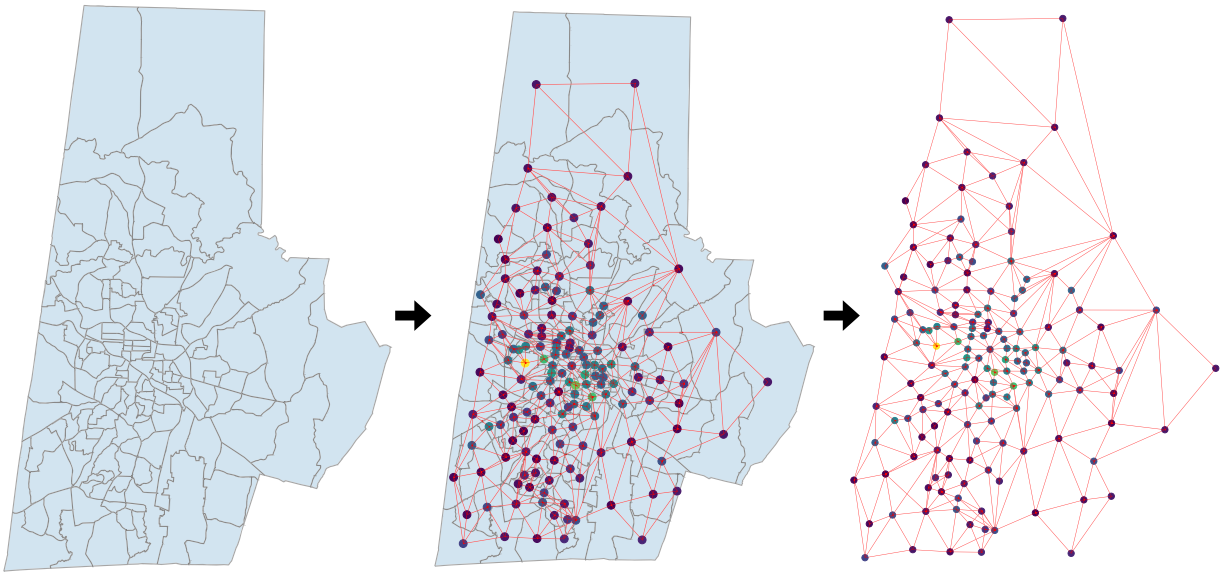


Figure 2. Illustration of the graph building process. Each block group is given its own node in the graph, and nodes are connected if they are neighbors to one another.

Statistical analysis

Our proposed method integrated spatial patterns with personalized information. First, we constructed time-varying graphs where each node represented a block group and each edge connected two neighboring block groups. Each year had a distinct graph that incorporated SDOH, climate factors and geographic patterns. Second, we predicted individual BMI using 1) a person's demographic information, and 2) the constructed graphs. This involved training a single graph neural network that works across all years to summarize information from the extended neighborhoods into a vector. We then combined this vector of summary features with each person's demographic features to predict BMI. The approach is described in detail below.

Building time-varying graphs

We first constructed a separate graph for each year t (ie, 2014-2022) to capture the impact of the extended neighborhood, with $t \in \{2014, 2015, \ldots, 2022\}$. To do this, we defined each block group j as a node, and we identified all contiguous block groups (ie, block groups that touched the index block group) to find "edges," as illustrated in Figure 2. We denoted all block groups in year t as $\boldsymbol{B}[t]$ and represented the j-th block group as $b_j[t]$ $(b_j[t] \in \boldsymbol{B}[t])$. Each block group contained features corresponding to that block group and year. The resulting graph included features of the node and edges. The edges were denoted $\boldsymbol{E}[t]$. This resulted in a graph $\boldsymbol{G}_t = \{\boldsymbol{B}[t], \boldsymbol{E}[t]\}$ containing $|\boldsymbol{B}[t]|$ nodes and $|\boldsymbol{E}[t]|$ edges for year t.

Some predictors were missing for certain block groups in specific years, with the overall missing values accounting for approximately 8% of the entire dataset. To address this, for each year, missing feature values were imputed by replacing them with the arithmetic mean of non-missing values from neighboring block groups. This process was applied iteratively (twice).

Individual outcome prediction framework

We generated individual-level predictions of BMI by integrating neighborhood-level contextual factors with individual demographics. Specifically, we employed a GNN with shared parameters to capture neighborhood effects across all years, enabling it to generalize across different graphs and temporal SDOH and climate patterns. This generalization is necessary as the precise location and number of block groups and census tracts can vary over years, meaning that the derived graph can be different for different years. By shared parameters, we mean that the same GNN parameters are used in each year that the GNN is applied, meaning that the predictions in different years can effectively share information on how best to learn the GNN.

Graph neural network modeling

We employed a unified graph convolutional neural network architecture[28] that can process all annual graphs, $\{\boldsymbol{G}_{2014}, \boldsymbol{G}_{2015}, \ldots, \boldsymbol{G}_{2022}\}$, using shared parameters, as shown in Figure 3b. For any single year, the network operated on that year's graph through the following steps:

Graph initialization

For each year's graph, we defined the node (ie, block group) predictors based on predictors from that specific block group and that specific year (eg, median household income for the block group with FIPS code 370630011001 in 2021).

Shared feature learning

We implemented a 2-layer graph convolutional network using shared weights across all years. To briefly summarize, in each layer each node (block group) summarizes its information along with the information of its neighbors. This happens twice, meaning that the output on each node can capture a summary of the information of the input features, the features of its neighbors, and the features of the neighbors of its neighbors, in what is often referred to as the 2-neighborhood. The extent of neighbors that is used is a tunable parameter and could be varied. This graph neural network performed iteratively passing messages (eg, information/summaries about their input features and prior messages passed from neighbors) between neighboring nodes to incorporate nearby information, and was defined[28] for each layer as:

$$\boldsymbol{h}_{b_j}^{l} = \left(\sum_{k \in \boldsymbol{N}_{b_j}[t] \cup \{\boldsymbol{b}_j[t]\}} \frac{1}{\sqrt{\hat{d}_{b_j} \hat{d}_k}} h_k^{(l-1)} W^{l} \right).$$

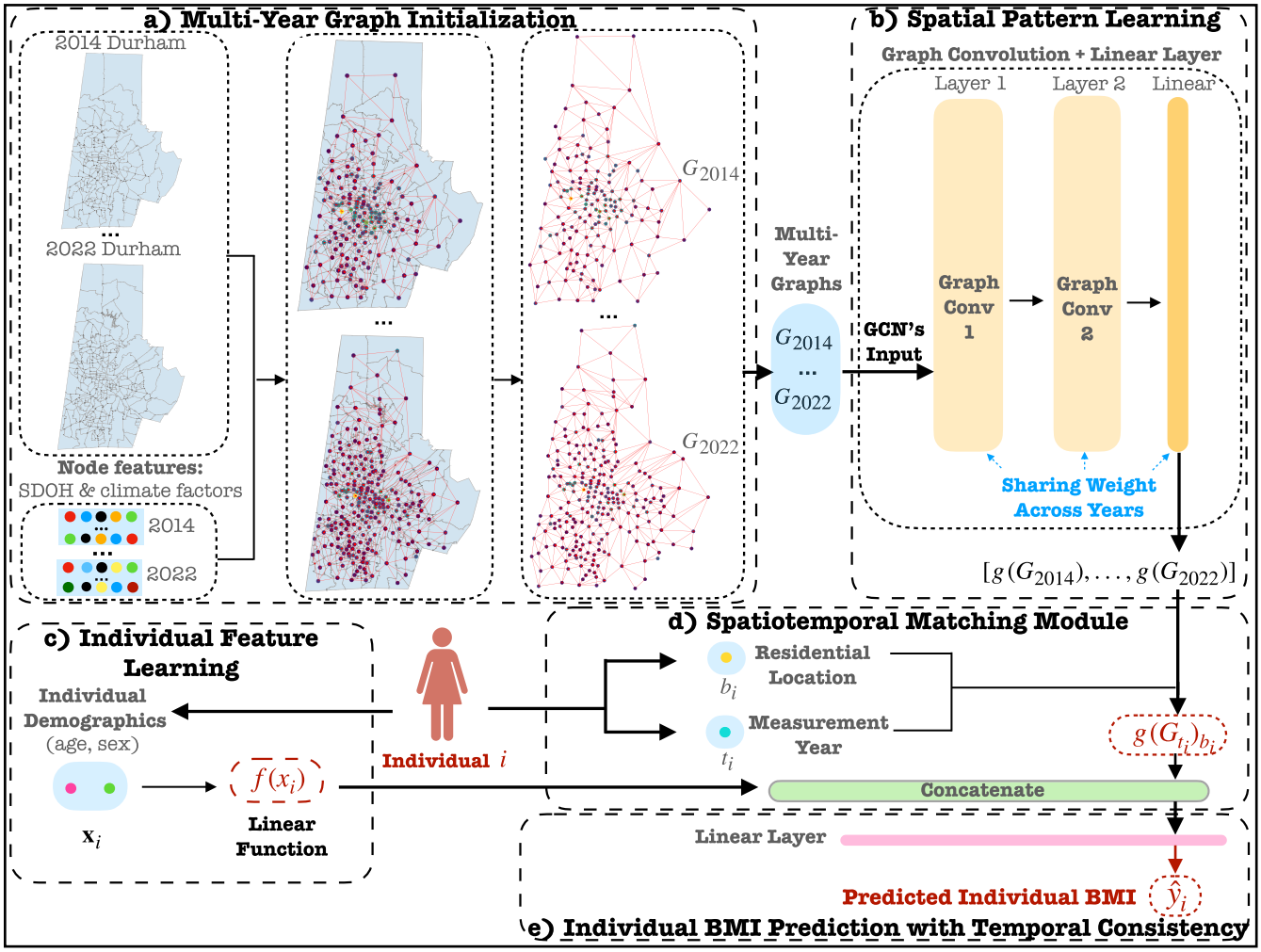


Figure 3. Schematic of prediction framework. a) Year-by-year graph construction spanning 2014 to 2022, as more fully described in Figure 2; b) spatial pattern learning via a graph convolutional network that captures index-level and neighboring area features, employing weight sharing across years; c) fully connected network for individual demographic information extraction; d)) spatiotemporal matching module for retrieving block group–level representation using the individual's residential location and BMI measurement year; e) individualized BMI prediction using demographics, geographic location, and block group–level data from the measurement year. BMI: Body mass index; GNN: Graph neural network.

where W^l is trainable weight matrix at layer l shared by all years; N_{b_j} [t] denotes neighbors of j-th block group in year t's graph; d_{b_j} is the degree of node b_j including a self-loop (eg, the number of block groups, or nodes, the j-th block group is connected to, counting a connection to itself); σ denotes a non-linear activation function, which is here we used the common Rectified Linear Unit (ReLU), which is defined as $ReLU(x) = 1_{x>0} \times x$. This definition means that the function returns 0 if x is less then 0, and x otherwise. This nonlinearity is necessary in deep networks to increase the capacity or flexibility of the network.

We denote the entire process in this section using function g. Thus, for year t, the output of the GNN was given as $g(\mathbf{G}_t)$, with block group b_j at time t given as $g(\mathbf{G}_t)_{b_j}$. Thus, the final embeddings for 2014 to 2022 of our graph neural network model were $\left[g(\mathbf{G}_{2014}), \ldots, g(\mathbf{G}_{2022})\right]$, that will then be used in predictions.

The proposed neighborhood-level learning architecture enables each node to refine its features by considering information from neighboring nodes, thereby capturing greater information about the extended neighborhood. This approach offers several key advantages: (1) it supports graphs that vary across years; (2) it adapts to year-specific adjacency structures; and (3) it maintains a consistent feature space across temporal domains.

Individual outcome prediction

As outlined in Figure 3, our framework provides patient-level BMI predictions for each individual i based on their demographic characteristics, while also considering predictors within their index and expanded neighborhoods. This process is shown in Figure 3c, where the individual level information $\mathbf{x}_i$ (ie, age and sex) measured in $year_i$ is first summarized by a simple neural network $f(x_i)$, and that information is combined with the summary values from the GNN for the block group that represents the information present in the extended neighborhood, $g(G_{t_i})_{b_i}$, meaning that we use the graph from the corresponding year and location of individual i. These two pieces of information are combined (Figure 3d) to make the final prediction y_i (Figure 3e), as

$$\hat{y}_i = \mathbf{W}\left(f(\mathbf{x}_i) \parallel g(\mathbf{G}_{t_i})_{b_i}\right) + c.$$

where $\parallel$ denotes concatenation of vectors, and $\mathbf{W}$ and c denote parameters of a linear transformation. In other words, this is the

same mathematical form as a linear regression given the two sets of summaries from the personal and location information.

During model training, the unified architecture learned from heterogeneous graphs across each year (2014-2022) to enable end-to-end training, meaning that the parameters of the GNN, the network $f(\cdot)$, W, and c are learned simultaneously to better interact with each other, using real BMI values as ground truth, with two key innovations:

- **Temporal generalization capability**: The shared-weight GNN processes graphs from any year without retraining
- **Dynamic topology handling**: Automatic adaptation to yearly changes in block group boundaries/connections that can happen due to updates from the decennial census.

Model training and selection

We generated a test set by randomly withholding 20% of the individuals prior to training. We used 5-fold cross-validation to estimate performance on the rest of the data. During training, we used the Adam optimizer[29] with a learning rate decay strategy[29] and mean squared error (MSE) loss. We optimized hyperparameters using Ax optimization.[30] These were, the learning rate of the Adam optimizer (range [1e-5, 0.4]); weight decay [5e-5, 5e-3]; learning rate decay step size [20, 50] and γ [0.6, 1.0]. We conducted 30 runs to identify the optimized hyperparameters for each experiment.

Model evaluation

Similar to other studies,[31,32] we evaluated inter-individual coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE) on a validation set. We compared to multiple baseline models, including benchmarking using the average BMI value. To evaluate the contribution of key model components, we conducted three model comparison experiments, each designed to isolate and assess the impact of specific features or structures within our framework. We chose to use two machine learning methods that do not capture extended neighborhoods as the baseline approaches to isolate the gain from capturing extended neighborhoods while keeping the nonlinear capabilities of the machine learning approaches, along with a penalized linear regression approach that cannot capture nonlinear relationships. Specifically, these approaches are: 1) LocalFCN: We replaced the GNN with a Fully Connected Neural Network (FCN) to only use the index block group, meaning that this uses a standard deep learning approach that does not capture the extended neighborhood; 2) We trained an Explainable Boosting Machine (EBM, ie, an explainable generalized additive model)[33] using characteristics in index residential block group and individual information, which likewise will not capture extended neighborhoods, but has the benefit that feature importance and nonlinear relationships can be easily visualized due to its "glassbox" model structure; 3) Linear regression with L2 penalty using characteristics in index residential block group and individual information. To summarize: the first two comparison models demonstrated the value of incorporating extended neighborhood information, while the third highlighted the predictive improvement gained through added non-linearity (as compared to traditional linear models). To assess the impact of child opportunity index (COI)—a metric partially derived from other features and partly incorporating unique external information[34]—we trained our model with and without COI. All analyses were conducted in Python3, version 3.10.4[35] with the packages, NetworkX, version 3.1[36] to create the graph, and PyTorch, version 2.0.1[37] to build the deep learning framework.

Results

Final cohort

The final study cohort included 12,226 children with a mean age of 11.5 years, of whom 53% were female, 45% were Black, and 25% were Hispanic (Table 1). When comparing characteristics among children above and below 12 years of age, older children were more likely to be female (55% vs 50%), Black (48% vs. 41%), and less likely to be Hispanic (21% vs. 27%), compared to children less than 12 years of age.

The overall block group-level mean values of social, environmental and climate predictors are presented in Table 2. Year-specific summary statistics across block groups from 2014 to 2022 are available in supplementary material (Tables S1-S9). These tables highlight substantial heterogeneity in socio-environmental characteristics across block groups, including in the yearly distributions of the COI, median household income, and the percentage of the population living below the poverty level.

Prediction performance and comparison models

We assessed the contribution of neighborhood social, environmental and climate data to BMI prediction and further evaluated the specific influence of including COI by comparing model performance with and without COI in the feature set (Table 3).

As shown in Table 3, our proposed method outperformed all comparison models, achieving the highest predictive performance both with and without COI. Of note, our model, which does account for autocorrelation between block groups, outperformed the Explainable Boosting Machine and the LocalFCN, two approaches that do not use autocorrelation between block groups for its predictions. These results underscored the importance of incorporating data from expanded neighborhoods for both age groups, regardless of whether COI is included. Results from the linear regression showed the importance of accounting for nonlinearity in our framework.

To further investigate the enhanced performance resulting from our model's capability to integrate data from expanded neighborhoods, we conducted comparative experiments using a modified test set, where demographics for all individuals were fixed. This ensured that only block-group level characteristics (SDOH and climate data) contributed to the inter-individual variation in BMI prediction. We split the visualizations in 2020, as block group boundaries and definitions were updated in that year, affecting data consistency across the 2014-2022 period. We used the feature set excluding COI for this analysis to avoid potential redundancy, as COI is a composite index partially derived from multiple block-group–level variables already included in the model. This approach allowed us to more directly assess the contribution of individual neighborhood features and the added value of incorporating spatial context.

We compared our model to LocalFCN by visualizing block group–level average BMI predictions (Figure 4) and calculating inter-individual R^2 on a modified test set. LocalFCN achieved R^2 values of 0.057 (ages 6-11) and 0.070 (ages 12-18), while our model improved performance to 0.088 and 0.132, respectively. These results suggest our model better captures inter-individual BMI variation from neighborhood-level context, especially among older children. Visualizations further support this, showing higher average BMI and greater spatial variability in the 12-18 age group. Our model's predictions align more closely with observed values than LocalFCN, highlighting the value of incorporating broader neighborhood information.

Table 1. Summary statistics on the baseline characteristics of the study population with standard deviations.

Characteristic	6-11 years old (n = 6929)	12-18 years old (n = 5397)	Total (n = 12226)
Age (years), mean ± SD	8.9 ± 1.7	14.6 ± 1.7	11.5 ± 3.6
BMI (kg/m^2), mean ± SD	20.6 ± 5.0	26.3 ± 6.9	23.1 ± 6.6
Female [n (%)]	3485 (50.3%)	2978 (55.2%)	6463 (52.9 %)
Race [n (%)]			
Black	2839 (40.9%)	2617 (48.4%)	5456 (44.6 %)
Caucasian/White	2321 (33.5%)	1568 (29.1%)	3889 (31.8 %)
Asian	189 (2.7%)	140 (2.6%)	329 (2.7 %)
Other	985 (14.2%)	541 (10.0%)	1562 (12.8 %)
More than one race	81 (1.2%)	85 (1.6%)	166 (1.4 %)
Missing[a]	511 (7.4%)	308 (5.7%)	819 (6.7 %)
Ethnic group [n (%)]			
Hispanic or Latino	1895 (27.3%)	1115 (20.6%)	3010 (24.6 %)
Not Hispanic or Latino	4693 (67.7%)	3989 (73.9%)	8682 (71.0 %)
Missing[a]	338 (4.9%)	190 (3.5%)	528 (4.3 %)

Abbreviations: BMI, Body Mass Index; SD, standard deviation.
[a]Corresponding characteristic is not included or not reported.

Significance tests

We validated the model by testing the statistical significance of incorporating SDOH and climate data. Specifically, for both age groups, we conducted analyses using our proposed model on different feature sets: 1) demographic information only (denoted as demo); 2) demographic information and SDOH (demo+SDOH); 3) demographic information and climate data (demo+temp); and 4) all features combined (demo+SDOH+temp). After collecting the resulting R^2 from 5 runs, we conducted t-tests and calculated the p-values for different pairs of feature sets.

Table 4 presents R^2 for both age groups across different predictor sets, with p-values indicating whether the addition of socioeconomic and environmental variables significantly improves BMI prediction. The results demonstrate that socioeconomic factors substantially improve model performance beyond demographics alone for both age groups ($P < 0.001$). In contrast, the contribution of temperature data is comparatively modest. Adding temperature alone to demographic features does not yield a statistically significant improvement ($P = 0.09$ for ages 6-11; $P = 0.11$ for ages 12-18). Even when SDOH features are already included, the incremental benefit of adding temperature remains limited—statistically non-significant for younger children ($P = 0.44$) and only marginally significant for adolescents ($P = 0.03$). These results suggest that while temperature may provide some additional predictive value, particularly for older children, the primary gains in BMI prediction are driven by neighborhood socioeconomic characteristics.

Table 2. Overall mean and standard deviation derived from the annual means for block group-level social, environmental and climate predictors across years.

Predictors	Value (mean ± standard deviation (SD))
Socioeconomic	
Child Opportunity Index (scale: 0-100)[a]	51.5 ± 4.4
Median household income ($)	64 653.5 ± 9817.2
Unemployment (%)	6.0 ± 1.6
Residents without vehicle (%)	9.2 ± 1.4
Households with high rent burden (%)	50.6 ± 2.4
Individuals living at or below poverty level (%)	16.9 ± 2.7
Households with more than one person per room (%)	2.8 ± 0.2
Built environment	
Tree coverage (%)	17.8 ± 0.5
National Walkability Index[b]	10.6 ± 4.0
Number of tobacco stores	1.2 ± 0.3
Number of liquor stores	1.2 ± 0.3
Number of grocery stores	0.3 ± 0.1
Number of recreation centers	0.5 ± 0.2
Number of convenience stores	0.3 ± 0.1
Number of religious institutions	1.8 ± 0.5
Number of fast food restaurants	0.4 ± 1.1
Landsat Normalized Difference Vegetation Index (NDVI)[c]	0.3 ± 0.0
Climate	
Temperature	26.8 ± 0.0

Abbreviation: SD, standard deviation over block groups.
All statistics are calculated using block group level data of block groups in Durham County.
[a]Diverse Data Kids (https://www.diversitydatakids.org/child-opportunity-index).
[b]USEPA (https://www.epa.gov/smartgrowth/national-walkability-index-user-guide-and-methodology).
[c]USGS (https://www.usgs.gov/landsat-missions/landsat-normalized-difference-vegetation-index).

Table 3. Model performance comparison for BMI value prediction.

Model	E (R^2)		E(MAE)		E(RMSE)	
	w COI	wo COI	w COI	wo COI	w COI	wo COI
6-11 years old						
Benchmark	0		3.969		4.944	
Linear model	0.166	0.169	3.550	3.615	4.495	4.603
EBM[33]	0.188	0.195	3.467	3.519	4.447	4.509
LocalFCN	0.228	0.227	3.368	3.376	4.384	4.395
Our model	0.234	0.230	3.352	3.369	4.370	4.386
12-18 years old						
Benchmark	0		5.349		6.601	
Linear model	0.078	0.082	5.220	5.299	6.514	6.555
EBM[33]	0.116	0.107	5.135	5.143	6.444	6.447
LocalFCN	0.139	0.131	4.991	5.031	6.366	6.449
Our model	0.147	0.134	4.980	5.018	6.342	6.389

Benchmark: model using average BMI as predicted value for all individuals; Linear Model: Linear Regression model using L2 penalty; EBM: an explainable generalized additive model; LocalFCN: uses index and neighborhood characteristics for predicting individual BMI; Our model: the proposed individual BMI prediction framework; MAE: mean absolute error; R^2: coefficients of determination for unseen independent data; RMSE: root mean square error.

Model-identified predictors and their importance

To understand which variables contributed most to BMI prediction, we examined feature importance scores, which are quantitative measures of how much each predictor influenced the model's output. Feature importance was calculated by EBM for each block group level predictor (Figure 5) and provided the full visualization of feature importance in supplementary material (Figure S2). While this framework does not perform as well as the full model, it has a tractable computation methodology to visualize the importance of each feature through established code packages. We excluded COI to clearly identify the contribution of specific social, environmental and climate predictors, which may otherwise be masked by the composite nature of COI. We also excluded individual demographics to focus specifically on block group-level information.

Figure 5 shows that block group-level SDOH and climate factors are more influential in adolescence than early childhood, consistent with our findings where socioeconomic and environmental factors explained more BMI variation in older children. Socioeconomic variables consistently rank above built environment and climate features. Built environment indicators like tree coverage and number of religious institutions also contribute meaningfully, with tree coverage ranking highest for ages 12-18. Temperature shows moderate importance, with scores of 0.066 (ages 6-11) and 0.117 (ages 12-18). Overall, while socioeconomic conditions remain the most influential drivers of obesity, built environment and climate features significantly contribute to predictions.

Table 4. Impact of using different predictor sets on prediction characteristics.

Significant Tests		
	p-value (on R^2)	
	6-11 years old	12-18 years old
demo vs. demo +temp	0.09	0.11
demo +SDOH vs. demo +SDOH+temp	0.44	0.03
demo vs. demo+SDOH	0.00	0.00
demo+temp vs. demo+SDOH+temp	0.00	0.00
demo vs. demo+SDOH+temp	0.00	0.00

All statistical values are calculated using results from our proposed individual BMI prediction framework. P-values are calculated based on all R^2 and MSE after 5 runs. demo: individual demographic features (age and sex); SDOH: Social and environmental Determinants of Health (Child Opportunity Index, median household income, unemployment rate, percentage of residents without vehicle, percentage of household with high rent burden, percent of individual living at or below poverty level, percentage of household with more than one person per room, tree coverage, National Walkability Index, number of tobacco stores, number of liquor store, number of grocery stores, number of recreation centers, number of convenience stores, number of religious institutions, number of fast food restaurants, Landsat Normalized Difference Vegetation Index); temp: summer evening temperature in Durham County; MSE: mean square error; R^2: coefficients of determination for unseen independent data.

Discussion

We developed a novel individual BMI prediction framework based on a GNN to include SDOH and climate information that may inform strategies to curtail childhood obesity. This framework shows that extended neighborhood information improves predictions compared to only information directly proximal to patients.[38,39] Furthermore, this study underscores the importance of considering socio-environmental factors in a broader geographic area from where a child lives when predicting health measures such as BMI. There are multiple novel aspects of the current study outlined below.

First, a major methodologic advance of our work is the use of GNNs to account for autocorrelation. Increasingly, GNN-based models are used to predict individual health outcomes.[40] However, work to date has not blended shared information over nodes (neighborhoods) with personal information. Specifically, prior studies that use graph structures for individual level disease prediction treat each patient or disease as a node,[40-43] which is incapable of directly learning shared relationships over space. To our knowledge, our study is the first to use GNNs to relate broader environmental features (eg, tree coverage, number of fast-food restaurants) to health by capturing the impact of extended neighborhoods on individuals.

Second, our GNN effectively models complex relationships and dependencies inherent in graph-structured data. Only a few studies have incorporated environmental variables.[32,44-47] Among these, the majority have focused on neighborhood environments directly proximal to where a person lives, whereas we show there is greater information captured by expanded neighborhoods. In

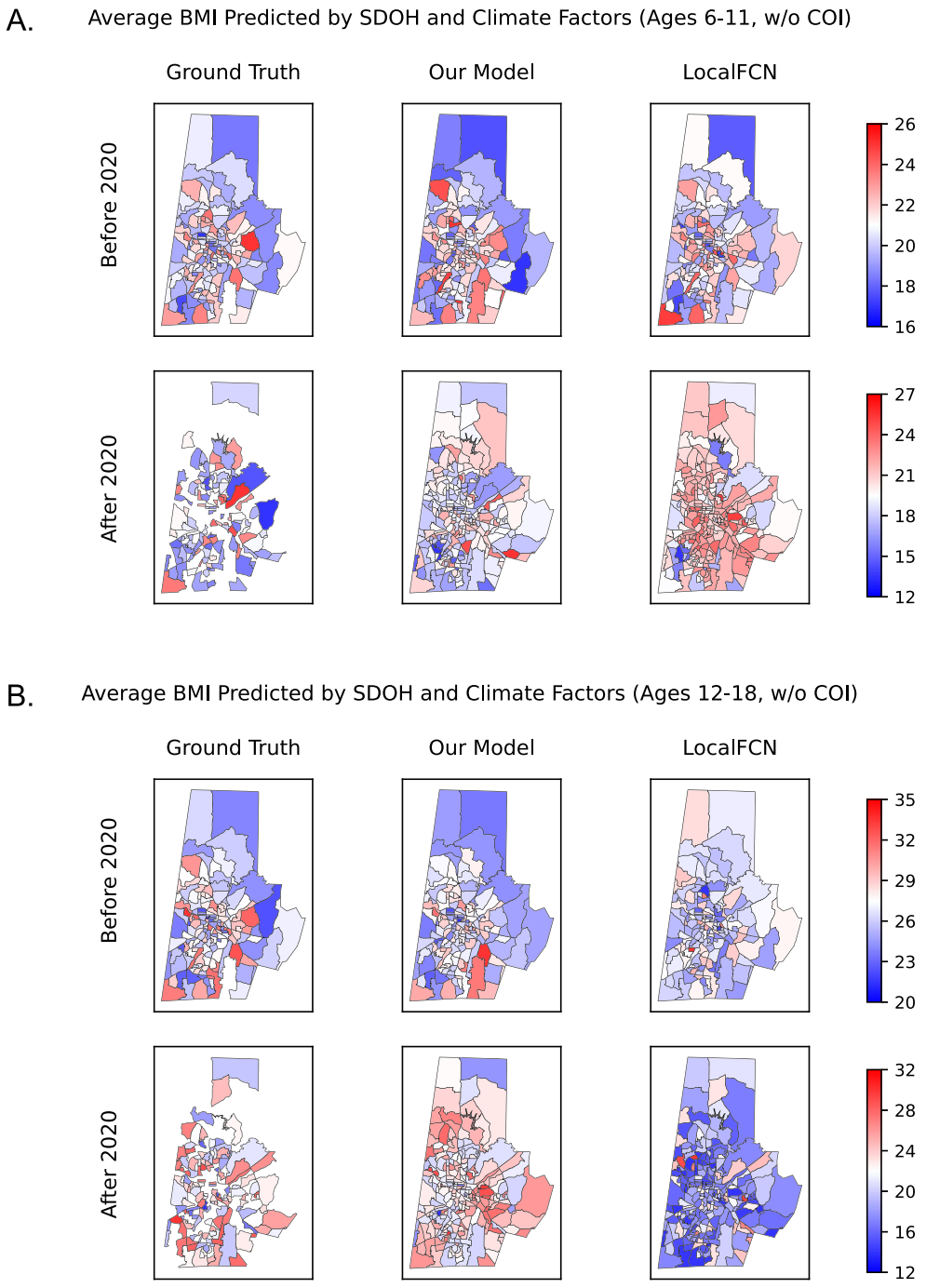


Figure 4. Block group level average BMI (kg/m^2) prediction using only SDOH and climate data for kids in both age groups. A) 6-11 years old group; B) 12-18 years old group. Visualizations are split at the year 2020 due to changes in block group definitions starting that year. Prediction results come from test (held-out) data. Each color bar represents the value of the average BMI (kg/m^2) prediction in their corresponding age group. Our full proposed model better captures the spatial variation in Durham County compared to only using information from an individual block group. Our model: The proposed individual BMI prediction framework; LocalFCN: The model which only uses index and neighborhood characteristics for predicting individual BMI; BMI: Body mass index.

addition, we extended the use of GNN by building a framework that incorporates personal and spatial factors and can be applied to dynamically changing graphs over years. This is critical in our study as the number of individuals within each node varies, resulting in different output dimensionality for each node. To overcome this challenge, our method simultaneously learns at a "sub-node" level, analyzing individual elements within each node concurrently. This approach addresses the shortcomings of traditional GNNs in handling variable output dimensions.

Third, our method of incorporating neighborhood-level temperature into the model is unique. Data products that quantify temperature, such as the ERA5-Land reanalysis dataset, provide information at too coarse a resolution (9 kilometers) to be useful at the block group level (as done in our study).[48] Moreover, the ERA5-Land dataset has been found to underestimate the intensity of urban heat.[49] Therefore, we opted to use a model of evening temperature estimates at high resolution so that we could capture the influence of the urban heat island effect when it tends to be

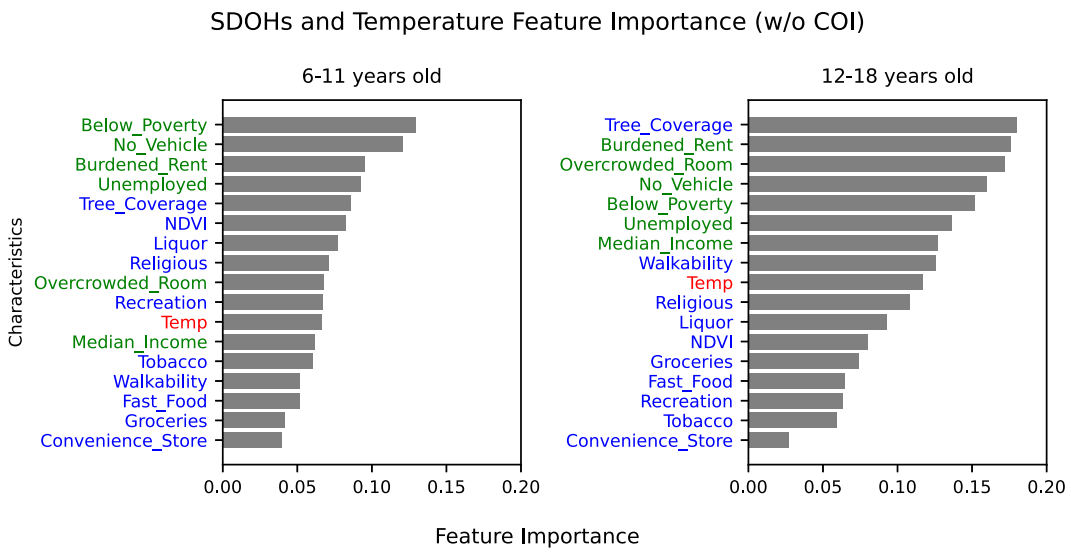


Figure 5. SDOH and temperature (climate data) feature importance after excluding COI. The x-axis indicates the estimated feature importance for each predictor; the y-axis indicates the name of each corresponding predictor. NDVI: LandSat normalized difference vegetation index.

most pronounced (ie, the urban–rural temperature difference is at its highest). We note that the temperature model is based on a snapshot of evening temperature and thus does not completely characterize urban heat island variability. However, because the data used in the model was collected under typical summer conditions, the resulting estimates serve as an indicator of average spatial temperature variation under peak impacts.

Fourth, our study simultaneously considers neighborhood sociodemographic, built environment, climate, and individual factors to predict individual BMI. This advances prior research that supports an association between neighborhood socio-environmental factors and obesity.[50-53] These studies have shown that children living in neighborhoods with poor housing and limited access to sidewalks, parks, and recreation centers, have higher odds of being overweight or obese compared to those in more favorable environments.[50] The physical neighborhood characteristics, such as the presence and distance to parks, highways, green streets, access to health food, and diverse housing types, correlate with variations in obesity prevalence across different neighborhoods.[50,53] Importantly, the risk of obesity increases with age, with younger children less sensitive to neighborhood environments than older ones.[53] Our study in Durham County aligned with these findings; SDOH and climate data explained 13.2% of the interindividual variance for children aged 12-18 and 8.8% for those aged 6-11. In addition, previous studies often only consider obesity prediction as a community-level outcome[50-60] rather than individual-level obesity prediction. Although recent methods target individual obesity prediction,[61-65] they primarily emphasize lifestyle, genetic, demographic, and clinical factors, and are unable to assess broader neighborhood influences.

There are some limitations to our study. First, our method uses a single BMI measurement at one point in time and does not account for length of exposure to neighborhood influences, which may influence the effect of SDOH and climate exposures[66-68]; longer-term residents may be more affected by local conditions than recent movers. Additionally, our findings may be less generalizable in areas with low socioeconomic and environmental variability, such as uniformly rural counties, where expanded neighborhood context may add limited predictive value. Finally, while our model outperforms all baseline approaches, the R^2 values may appear modest in absolute terms. This is expected given the high inter-individual variability in BMI and the influence of lifestyle, diet, physical activity, and genetics, which are known predictors of obesity. Despite these constraints, our approach captured more variance than established models and revealed clear spatial patterns aligned with known socio-environmental disparities.

Future research should include broader geographic areas and larger populations to better assess the impact of SDOH and climate on BMI. Considering the length of time spent at current and prior addresses of individuals may clarify how cumulative exposures influence outcomes. Finally, identifying causal pathways could inform targeted obesity prevention strategies.

Conclusion

In conclusion, we observed statistically significant improvements in prediction performance of a model predicting child and adolescent BMI when using a graph neural network (GNN), suggesting that accounting for spatial autocorrelation in socio-environmental factors is important in childhood obesity prediction. While prior studies have accounted for spatial autocorrelation when predicting clinical outcomes, they are not able to effectively model complex relationships and dependencies inherent in graph-structured data. Our study addresses these limitations by integrating individual-level data with both proximal and expanded neighborhood-level contextual factors using a GNN framework. The results highlight the importance of broader neighborhood socioeconomic and environmental exposures beyond a child's immediate surroundings in shaping BMI. These findings inform public policy, urban planning, and community development approaches to address the childhood obesity epidemic, while also offering a novel, generalizable modeling for other clinical conditions.

Author contributions

Keyu Li (Conceptualization [equal], Data curation [lead], Formal analysis [lead], Investigation [supporting], Methodology [lead],

Project administration [equal], Software [lead], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [lead]), Charles Wood (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Supervision [equal], Visualization [supporting], Writing—review & editing [supporting]), Liz Nichols (Data curation [equal], Formal analysis [supporting]), Zachary Calhoun (Data curation [equal], Writing—original draft [supporting], Writing—review & editing [supporting]), Nrupen Bhavsar (Conceptualization [lead], Data curation [lead], Formal analysis [lead], Funding acquisition [lead], Investigation [lead], Methodology [supporting], Project administration [lead], Resources [lead], Supervision [lead], Validation [equal], Visualization [supporting], Writing—original draft [equal], Writing—review & editing [lead]), and David Carlson (Conceptualization [lead], Data curation [supporting], Formal analysis [lead], Funding acquisition [lead], Investigation [lead], Methodology [lead], Project administration [lead], Resources [lead], Supervision [lead], Validation [lead], Visualization [supporting], Writing—review & editing [lead])

Supplementary material

Supplementary material is available at *AJE Advances: Research in Epidemiology* online.

Funding

This study was supported by the MEDx program through the Duke University. K.L. was supported by the MEDx program, N.A.B. was supported by NIH HL140146 and UL1TR002553, C.W. was supported by NIH K23HD107157.

Conflicts of interest

The authors declare no conflicts of interests.

Disclaimer

The views expressed in this manuscript are the authors' views and do not necessarily reflect the official policies of the NIH.

Data availability

The source of socioeconomic, built environment and climate data used in this work are either available on the Social, Environmental, and Equity Drivers (SEED) Health Atlas: https://sdoh.duhs.duke.edu/atlas or upon request. The individual demographic and health outcome data used in this study are available on reasonable request and following institutional policies from the corresponding authors, N.A.B. and D.C.

References

1. Geserick M, Vogel M, Gausche R, et al. Acceleration of BMI in early childhood and risk of sustained obesity. *N Engl J Med*. 2018;379(14):1303-1312. https://doi.org/10.1056/NEJMoa1803527
2. Ward ZJ, Long MW, Resch SC, et al. Simulation of growth trajectories of childhood obesity into adulthood. *N Engl J Med*. 2017;377(22):2145-2153. https://doi.org/10.1056/NEJMoa1703860
3. Whitaker RC, Wright JA, Pepe MS, et al. Predicting obesity in young adulthood from childhood and parental obesity. *N Engl J Med*. 1997;337(13):869-873. https://doi.org/10.1056/NEJM199709253371301
4. Ogden CL, Fryar CD, Martin CB, et al. Trends in obesity prevalence by race and Hispanic origin—1999-2000 to 2017-2018. *JAMA*. 2020;324(12):1208-1210. https://doi.org/10.1001/jama.2020.14590
5. Simmonds M, Llewellyn A, Owen CG, et al. Predicting adult obesity from childhood obesity: a systematic review and meta-analysis. *Obes Rev*. 2016;17(2):95-107. https://doi.org/10.1111/obr.12334
6. Goodman E. The role of socioeconomic status gradients in explaining differences in US adolescents' health. *Am J Public Health*. 1999;89(10):1522-1528. https://doi.org/10.2105/AJPH.89.10.1522
7. Swinburn B, Egger G, Raza F. Dissecting obesogenic environments: the development and application of a framework for identifying and prioritizing environmental interventions for obesity. *Prev Med*. 1999;29(6):563-570. https://doi.org/10.1006/pmed.1999.0585
8. Ziauddeen N, Roderick PJ, Macklon NS, et al. Predicting childhood overweight and obesity using maternal and early life risk factors: a systematic review. *Obes Rev*. 2018;19(3):302-312. https://doi.org/10.1111/obr.12640
9. Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. https://doi.org/10.1126/science.aax2342
10. Liu P, De Sabbata S. Spatial Autocorrelation Analysis with Graph Convolutional Neural Network. *29th Annual GIS Research UK (GISRUK) Conference*. 2021; https://doi.org/10.5281/zenodo.4665834
11. Stolte, Merli MG, Hurst JH, et al. Using electronic health records to understand the population of local children captured in a large health system in Durham County, NC, USA, and implications for population health research. *Soc Sci Med*. 2022;296:114759. https://doi.org/10.1016/j.socscimed.2022.114759
12. Singh GK, Siahpush M, Kogan MD. Neighborhood socioeconomic conditions, built environments, and childhood obesity. *Health Aff*. 2010;29(3):503-512. https://doi.org/10.1377/hlthaff.2009.0730
13. Maharana A, Nsoesie EO. Use of deep learning to examine the association of the built environment with prevalence of neighborhood adult obesity. *JAMA Netw Open*. 2018;1(4):e181535. https://doi.org/10.1001/jamanetworkopen.2018.1535
14. Kim Y, Cubbin C, Oh S. A systematic review of neighbourhood economic context on child obesity and obesity-related behaviours. *Obes Rev*. 2019;20(3):420-431. https://doi.org/10.1111/obr.12792
15. Yang Y, Jiang Y, Xu Y, et al. A cross-sectional study of the influence of neighborhood environment on childhood overweight and obesity: variation by age, gender, and environment characteristics. *Prev Med*. 2018;108:23-28. https://doi.org/10.1016/j.ypmed.2017.12.021
16. Lanza K, Alcazar M, Hoelscher DM, et al. Effects of trees, gardens, and nature trails on heat index and child health: design and methods of the green schoolyards project. *BMC Public Health*. 2021;21:1-2. https://doi.org/10.1186/s12889-020-10128-2
17. Maddren CI, Dhamrait G, Ghogho M, et al. Parental perceptions of environmental factors on preschoolers' outdoor play in 19 low-income, middle-income, and high-income countries. *J Phys Act Health*. 2025;1(aop):1-1.
18. Durham Neighborhood Compass. Housing. Accessed February 10, 2025. https://compass.durhamnc.gov/en/

19. Data Axle. Reference Solutions. Accessed February 15, 2025. https://www.referenceusagov.com/
20. United States Environmental Protection Agency. National Walkability Index User Guide and Methodology. Accessed February 5, 2025. https://www.epa.gov/smartgrowth/national-walkability-index-user-guide-and-methodology
21. North Carolina State Climate Office. Urban heat islands. Accessed February 20, 2025. https://climate.ncsu.edu/research/uhi/
22. U.S. Geological Survey. Annual NLCD (National Land Cover Database)—The next generation of land cover mapping. Accessed February 20, 2025. https://pubs.usgs.gov/publication/fs20253001
23. Gorelick N, Hancher M, Dixon M, et al. Google earth engine: planetary-scale geospatial analysis for everyone. *Remote Sens Environ*. 2017;202:18-27. https://doi.org/10.1016/j.rse.2017.06.031
24. Zolotor A, Huang RW, Bhavsar NA, et al. Comparing social disadvantage indices in pediatric populations. *Pediatrics*. 2024;154(3):1-9. https://doi.org/10.1542/peds.2023-064463
25. Calhoun ZD, Willard F, Ge C, et al. Estimating the effects of vegetation and increased albedo on the urban heat island effect with spatial causal inference. *Sci Rep*. 2024;14(1):540. https://doi.org/10.1038/s41598-023-50981-w
26. U.S. Centers for Disease Control and Prevention. Healthy Weight, Nutrition, and Physical Activity: Body Mass Index (BMI). Accessed November 20, 2022. https://www.cdc.gov/healthyweight/assessing/bmi/index.html
27. United States Census Bureau. Census Mapping Files: TIGER/Line Shapefiles. Accessed February 8, 2025. https://www.census.gov/
28. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. *5th Int Conf Learn Represent (ICLR)*. 2017; https://doi.org/10.48550/arXiv.1609.02907
29. Kingma DP, Ba J. Adam: a method for stochastic optimization-arXiv preprint. 2014. arXiv:1412.6980
30. Facebook Open Source. Adaptive Experimentation Platform. Accessed October 12, 2023. https://ax.dev/
31. Gray JC, Schvey NA, Tanofsky-Kraff M. Demographic, psychological, behavioral, and cognitive correlates of BMI in youth: findings from the adolescent brain cognitive development (ABCD) study. *Psychol Med*. 2020;50(9):1539-1547. https://doi.org/10.1017/S0033291719001545
32. Singh B, Tawfik H. A machine learning approach for predicting weight gain risks in young adults. *10th Int Conf Dependable Syst Serv Technol (DESSERT)*. 2019; 231-234. https://doi.org/10.1109/DESSERT.2019.8770016
33. Interpret ML. Explainable Boosting Machine. Accessed April 5, 2023. https://interpret.ml/docs/ebm.html#lou2013accurate-ebm
34. Noelke C, McArdle N, Baek M, et al. Child Opportunity Index 2.0 technical documentation. Accessed November 9, 2023. https://www.diversitydatakids
35. Python Software Foundation. Python 3.10.4. Accessed July 2, 2023. https://www.python.org/downloads/release/python-3104/
36. NetworkX. Network 3.3. Accessed October 20, 2023. https://github.com/networkx/networkx
37. The Linux Foundation. Pytorch 2.0. Accessed October 8. https://pytorch.org/get-started/pytorch-2.0/
38. Siddiqui H, Rattani A, Woods NK, et al. A survey on machine and deep learning models for childhood and adolescent obesity. *IEEE Access*. 2021;9:157337-157360. https://doi.org/10.1109/ACCESS.2021.3131128
39. Colmenarejo G. Machine learning models to predict childhood and adolescent obesity: a review. *Nutrients*. 2020;12(8):2466. https://doi.org/10.3390/nu12082466
40. Lu H, Uddin S. Disease prediction using graph machine learning based on electronic health data: a review of approaches and trends. *Healthcare*. 2023;11(7):1031. https://doi.org/10.3390/healthcare11071031
41. Lu H, Uddin S, Hajati F, et al. A patient network-based machine learning model for disease prediction: the case of type 2 diabetes mellitus. *Appl Intell*. 2022;52(3):2411-2422. https://doi.org/10.1007/s10489-021-02533-w
42. Khan A, Uddin S, Srinivasan U. Chronic disease prediction using administrative data and graph theory: the case of type 2 diabetes. *Expert Syst Appl*. 2019;136:230-241. https://doi.org/10.1016/j.eswa.2019.05.048
43. Hossain ME, Uddin S, Khan A. Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes. *Expert Syst Appl*. 2021;164:113918. https://doi.org/10.1016/j.eswa.2020.113918
44. Nau C, Ellis H, Huang H, et al. Exploring the forest instead of the trees: an innovative method for defining obesogenic and obesoprotective environments. *Health Place*. 2015;35:136-146. https://doi.org/10.1016/j.healthplace.2015.08.002
45. Hinojosa AM, MacLeod KE, Balmes J, et al. Influence of school environments on childhood obesity in California. *Environ Res*. 2018;166:100-107. https://doi.org/10.1016/j.envres.2018.04.022
46. Van Hulst A, Roy-Gagnon MH, Gauvin L, et al. Identifying risk profiles for childhood obesity using recursive partitioning based on individual, familial, and neighborhood environment factors. *Int J Behav Nutrition Phys Activity*. 2015;12:1-9. https://doi.org/10.1186/s12966-015-0175-7
47. Wiechmann P, Lora K, Branscum P, Fu J. Identifying discriminative attributes to gain insights regarding child obesity in hispanic preschoolers using machine learning techniques. *IEEE 29th Int. Conf. Tools Artif. Intell. (ICTAI)*. 2017; https://doi.org/10.1109/ICTAI.2017.00014
48. Muñoz-Sabater J, Dutra E, Agustí-Panareda A, et al. ERA5-land: a state-of-the-art global reanalysis dataset for land applications. *Earth Syst Sci Data*. 2021;13:4349-4383. https://doi.org/10.5194/essd-13-4349-2021
49. Lee J, Dessler AE. Improved surface urban heat impact assessment using GOES satellite data: a comparative study with ERA-5. *Geophys Res Lett*. 2024;51(1):e2023GL107364.
50. Singh GK, Siahpush M, Kogan MD. Neighborhood socioeconomic conditions, built environments, and childhood obesity. *Health Aff*. 2010;29(3):503-512. https://doi.org/10.1377/hlthaff.2009.0730
51. Maharana A, Nsoesie EO. Use of deep learning to examine the association of the built environment with prevalence of neighborhood adult obesity. *JAMA Netw Open*. 2018;1(4):e181535. https://doi.org/10.1001/jamanetworkopen.2018.1535
52. Kim Y, Cubbin C, Oh S. A systematic review of neighbourhood economic context on child obesity and obesity-related behaviours. *Obes Rev*. 2019;20(3):420-431. https://doi.org/10.1111/obr.12792
53. Yang Y, Jiang Y, Xu Y, et al. A cross-sectional study of the influence of neighborhood environment on childhood overweight and obesity: variation by age, gender, and environment characteristics. *Prev Med*. 2018;108:23-28. https://doi.org/10.1016/j.ypmed.2017.12.021
54. Lotfata A, Georganos S, Kalogirou S, et al. Ecological associations between obesity prevalence and neighborhood determinants using spatial machine learning in Chicago, Illinois,

USA. *ISPRS Int J Geo Inf*. 2022;11(11):550. https://doi.org/10.3390/ijgi11110550
55. Sun Y, Wang S, Sun X. Estimating neighbourhood-level prevalence of adult obesity by socio-economic, behavioural and built environment factors in new York City. *Public Health*. 2020;186:57-62. https://doi.org/10.1016/j.puhe.2020.05.003
56. Papas MA, Alberg AJ, Ewing R, et al. The built environment and obesity. *Epidemiol Rev*. 2007;29(1):129-143. https://doi.org/10.1093/epirev/mxm009
57. Booth KM, Pinkston MM, Poston WS. Obesity and the built environment. *J Am Diet Assoc*. 2005;105(5):110-117. https://doi.org/10.1016/j.jada.2005.02.045
58. Lam TM, Vaartjes I, Grobbee DE, et al. Associations between the built environment and obesity: an umbrella review. *Int J Health Geogr*. 2021;20:1-24. https://doi.org/10.1186/s12942-021-00260-6
59. Howell NA, Booth GL. The weight of place: built environment correlates of obesity and diabetes. *Endocr Rev*. 2022;43(6):966-983. https://doi.org/10.1210/endrev/bnac005
60. Mohammed SH, Habtewold TD, Birhanu MM, et al. Neighbourhood socioeconomic status and overweight/obesity: a systematic review and meta-analysis of epidemiological studies. *BMJ Open*. 2019;9(11):e028238. https://doi.org/10.1136/bmjopen-2018-028238
61. Tabrizi SS, Sancar N. Prediction of body mass index: a comparative study of multiple linear regression, ANN and ANFIS models. *Procedia Comput Sci*. 2017;120:394-401. https://doi.org/10.1016/j.procs.2017.11.255
62. Pang X, Forrest CB, Lê-Scherban F, et al. Prediction of early childhood obesity with machine learning and electronic health record data. *Int J Med Inform*. 2021;150:104454. https://doi.org/10.1016/j.ijmedinf.2021.104454
63. Ferdowsy F, Rahi KS, Jabiullah MI, et al. A machine learning approach for obesity risk prediction. *Curr Res Behav Sci*. 2021;2:100053. https://doi.org/10.1016/j.crbeha.2021.100053
64. Dirik M. Application of machine learning techniques for obesity prediction: a comparative study. *J Complexity Health Sci*. 2023;6(2):16-34. https://doi.org/10.21595/chs.2023.23193
65. Du J, Yang S, Zeng Y, et al. Visualization obesity risk prediction system based on machine learning. *Sci Rep*. 2024;14(1):22424. https://doi.org/10.1038/s41598-024-73826-6
66. Ahmed AT, Quinn VP, Caan B, et al. Generational status and duration of residence predict diabetes prevalence among Latinos: the California Men's health study. *BMC Public Health*. 2009;9:1-1. https://doi.org/10.1186/1471-2458-9-392
67. Cho Y, Frisbie WP, Hummer RA, et al. Nativity, duration of residence, and the health of Hispanic adults in the United States 1. *Int Migr Rev*. 2004;38(1):184-211. https://doi.org/10.1111/j.1747-7379.2004.tb00193.x
68. Spring A. Short-and long-term impacts of neighborhood built environment on self-rated health of older adults. *Gerontologist*. 2018;58(1):36-46. https://doi.org/10.1093/geront/gnx119

AJE Advances: Research in Epidemiology, 2025, **1**, 1–12
https://doi.org/10.1093/ajeadv/uuaf011
Original Contribution