Beyond BMI: Analysis of Obesity's Social Gradient in the United States

Varsha Chhabria, Nikitha Thota, Sravya Sree Ugni
HINF 6400: Introduction to Health Data Analytics
Professor Dan Ries
Northeastern University

Abstract

Obesity in the United States remains a complex public health challenge affecting over 42% of adults and 20% of youth. This analysis investigates the systemic drivers of obesity by synthesizing findings across three objectives. First, an ANOVA of adult obesity (BRFSS, 2011-2023) identified age as the most dominant factor (Sum of Squares = 426.2), with prevalence peaking in the 45-54 age group. Second, various machine learning classification models were trained on lifestyle factors like Sleep Duration, Physical Activity and Stress to predict obesity class. Most critically, an analysis of childhood obesity (NHANES) revealed a stark, inverse dose-response relationship between family income and obesity prevalence ($p = 7.13e-4$), with disparities accelerating during the school-age years. A comparative synthesis demonstrates that while childhood risk is uniformly driven by poverty, the adult socioeconomic gradient is confounded by gender and structural factors like occupation. The findings demonstrate that obesity is a disease of inequality, imprinted by childhood socioeconomic status and compounded by adult structural environments. Effective solutions must therefore target the economic, social and environmental drivers of the epidemic.

I. Introduction and Background

Obesity has become one of the most persistent and complex public health challenges in the United States, affecting millions of individuals across all age groups. According to the Centers for Disease Control and Prevention (CDC), more than 42% of American adults and nearly 20% of children and adolescents (ages 2–19) are classified as obese. This means more than 100 million U.S. adults live with obesity, including over 22 million adults with severe obesity, defined as a body mass index (BMI) of 40 or higher (Centers for Disease Control and Prevention [CDC], 2024). The CDC also reports that 14.7 million children and adolescents (ages 2–19) are classified as obese, representing about 1 in 5 (Centers for Disease Control and Prevention [CDC], 2024). Rates have more than doubled in children and tripled in adolescents over the past three decades, transforming obesity from an individual health concern into a population-level epidemic (Sanyaolu et al., 2019).

The consequences of this epidemic extend far beyond simple metrics of weight. Adults and children with obesity face elevated risks for chronic comorbities including hypertension, diabetes, cardiovascular disease, sleep apnea, osteoarthritis, and certain cancers. The impact extends beyond temporary health effects, shaping long-term wellbeing across the life course. Up to 80% of adolescents with obesity remain obese as adults, increasing their lifetime risk of cardiovascular and metabolic complications (Sanyaolu et al., 2019). Childhood obesity carries significant psychological and social consequences, including higher rates of depression, anxiety, and bullying compared to healthy-weight peers (Stohl, 2023). These trajectories established early in life are reinforced by social and environmental conditions in adulthood.

At a population level, obesity and overweight are among the fastest-growing contributors to morbidity and mortality in the United States, responsible for approximately 335,000 deaths and 11.6 million disability-adjusted life-years (DALYs) in 2021 (GBD 2021 US Obesity Forecasting Collaborators, 2024).

Increasingly, public health research has emphasized that obesity cannot be understood solely because of individual choices. Rather, it reflects the complex interaction between biological, behavioral, and social

environments. People's opportunities to eat nutritious food, engage in physical activity, and maintain a healthy weight are fundamentally shaped by the environments in which they are born. This relationship is best captured through the framework of Social Determinants of Health (SDOH**),** the nonmedical factors that influence health outcomes, including economic stability, education access and quality, neighborhood and built environment, social and community context, access to healthcare and food security (Randolph & Stephens, 2024). Inequities within these domains contribute to the uneven distribution of obesity and its related chronic diseases across populations.

Social, Environmental, and Behavioral Determinants

Social determinants such as income, education, and occupation influence access to nutritious food, healthcare, and opportunities for physical activity. Research consistently shows that individuals with lower socioeconomic status (SES) experience higher rates of obesity due to limited access to healthy foods, fewer recreational spaces, and greater exposure to low-cost, energy-dense foods. Longitudinal studies indicate that low SES across the life course, especially during childhood, contributes to increased obesity risk later in life. At the same time, obesity can reinforce economic disadvantage through discrimination and healthcare costs, demonstrating a bidirectional relationship (Autret & Bekelman, 2024). Educational attainment also plays a key role, as health literacy, nutrition awareness, and job stability are closely linked to lifestyle choices and dietary behaviors (Cohen et al., 2013).

The environment in which people live further reinforces or restricts healthy behaviors. Access to green spaces, parks, sidewalks, and recreational facilities encourages movement and social connection, while car dependent or high-traffic neighborhoods limit these opportunities. Conversely, food deserts, areas with limited access to affordable and nutritious foods and food swamps, characterized by a high density of fast-food outlets, are both associated with elevated obesity rates in children and adults (Polyzou & Polyzos, 2024).

Economic And Societal Burden

The economic and social costs of obesity extend far beyond healthcare expenditures, encompassing lost productivity, educational impacts, and even national readiness. In 2019, obesity-related medical care costs in the

United States were estimated at nearly $173 billion, with affected individuals incurring substantially higher expenses than those of normal weight (Centers for Disease Control and Prevention [CDC], 2022). More recent estimates place the direct health-care costs between $261 billion and $481 billion annually, ranking obesity among the nation's costliest chronic conditions (GBD 2021 US Obesity Forecasting Collaborators, 2024).

A less commonly discussed consequence of obesity is its impact on military readiness. More than one in three young adults are now considered too heavy to serve, and even among those who meet weight standards, many lack the physical stamina required for training. Among active-duty soldiers, obesity has been linked to a 33% higher injury risk (Centers for Disease Control and Prevention [CDC], 2022). Beyond physical health, obesity contributes to reduced educational attainment and lower quality of life, while weight reduction and bariatric surgery have been associated with improved well-being (Anekwe et al., 2020)

Project Need and Significance

The persistence and growth of obesity demonstrate a clear need for renewed public health attention and targeted research. While the issue has been widely studied, existing approaches have too often emphasized individual responsibility without fully addressing the systemic inequities that drive disparities in prevalence and outcomes. There is a pressing need to examine obesity not only as a clinical or behavioral concern but as a broader public health challenge shaped by social, environmental, and economic conditions (Kelly et al., 2025). This project focuses on these intersecting factors to highlight why current interventions remain insufficient and to underscore the urgency of developing comprehensive, equity-driven strategies capable of reducing disparities and improving health outcomes.

II. Research Objectives, Questions and Hypotheses

This report is structured around three core analytical objectives:

Objective 1: Analyze Temporal and Regional Trends of Obesity

Question: How has the prevalence of obesity evolved across the United States from 2011 to 2023, and what are the discernible regional patterns in these trends?

Hypothesis 1.1 Obesity prevalence in the U.S between 2011 and 2023.

H_0 (Null): There is no statistically significant change in the mean obesity prevalence between 2011 and 2023.

H_1 (Alternative): The mean obesity prevalence in 2023 is significantly different from the mean prevalence in 2011.

Analytical Methodology

- Statistical tests
    - A two-way ANOVA will be conducted to examine the effects of the year (2011–2023) and age group on mean obesity prevalence.
- Graphs
    - Line Graph will be used to compare mean obesity prevalence from 2011 - 2023 among each age group.
- Descriptive Statistics
    - Mean, standard deviation, minimum, and maximum were calculated for each age group per year.

Hypothesis 1.2 Obesity prevalence among different US states & regions.

H_0 (Null): There is no statistically significant difference in the mean obesity prevalence among the four designated U.S. Census regions.

H_1 (Alternative): At least one U.S. Census region exhibits a mean obesity prevalence that is statistically different from the others.

- **Statistical tests**

  - A two-way ANOVA will be conducted to examine the effects of obesity prevalence among different US regions over the years.

- **Graphs**
  - **Heat Map** — To display mean obesity prevalence by **state and year**, helping to identify high-obesity and low-obesity states and regional clustering trends.
  - **Regional Line Graph** — Show temporal trends in obesity prevalence across the four U.S. Census regions from 2011–2023.
  - **Bar Charts** showing the
    - Top 10 Most Obese States (2023)
    - Bottom 10 Least Obese States (2023)
- **Descriptive Analysis** will be calculated for each region to summarize mean obesity prevalence

Hypothesis 1.3 Obesity prevalence among different age groups in adults

H_0 (Null): There is no significant difference in mean obesity prevalence among the different age groups.

H_1 (Alternative): There is a significant difference in mean obesity prevalence among at least two of the age groups.

- **Statistical tests**
  - A one-way ANOVA will be conducted to examine whether mean obesity prevalence differs significantly across age groups.
- **Graphs**
  - A box plot will be used to visualize the distribution of obesity prevalence across different age groups from 2011–2023.
- **Descriptive Analysis**
  - The F-statistic and p-value will be used to underline the significance of age on obesity prevalence.

Objective 2: Investigate the Impact of Lifestyle Factors on Obesity

Question: What is the quantitative relationship between key lifestyle factors on obesity levels as measured by Body Mass Index (BMI)?

Hypothesis 2.1: Predictive Power of Lifestyle Variables

H_0: Sleep duration, stress, and physical activity do **not predict** BMI category better than random chance.

H_1: Sleep duration, stress, and physical activity can **accurately predict** BMI category.

- **Machine Learning & Statistical tests:**
  - A random forest will be used after splitting the data set into training and testing data with 80/20 splits.
- **Graphs**
  - A variable importance graph is plotted for Sleep Duration, Stress and Physical activity for comparison.
  - A sample random forest decision tree will be shown to understand the way the model is splitting the features.
- **Descriptive Analysis**
  - Predictions will be made on test data after training and accuracy will be measured of the random forest model.

Hypothesis 2.2: Feature Importance Comparison for BMI Prediction

H_0: All features contribute equally to BMI prediction.

H_1: Some features (like sleep duration or stress) have greater predictive importance than others.

- **Machine Learning & Statistical tests:**
  - A random forest will be built on all the features.

  - A chi-square test will be performed on the feature with high statistical significance on obesity levels.

- **Graphs**
  - A variable importance graph is plotted for all the features.
  - A sample random forest decision tree will be shown to understand the way the model is splitting the features.
- **Descriptive Analysis**
  - Predictions will be made on test data after training, and accuracy will be measured by the random forest model.

Hypothesis 2.3: Model Comparison Hypothesis

H_0: Simpler models (logistic regression) perform equally well as complex models (random forest, neural network).

H_1: Complex ML models outperform traditional models for BMI prediction.

- **Machine Learning & Statistical tests:**
  - Multiple machine learning models like logistic regression, random forests, and neural networks will be pitted against each by training / testing the same dataset.
- **Graphs**
  - Neural Network Graphs to visualize the underlying model will be shown.
  - Logistic regression Graphs to visualize the regression plane or line will be shown.
  - Random forest Tree Graph to visualize the splitting of various features will be shown.
- **Descriptive Analysis**
  - Compare accuracies and validate the hypothesis of each model.

Objective 3 - Child and Adolescent Obesity

Question. How has child/adolescent obesity prevalence varied in the U.S. over time and across key subgroups (sex, age bands, race/ethnicity, and income/poverty)?

Hypothesis 3.1 Obesity prevalence among different age groups in children

Hypotheses (Child)- Temporal Trend

H_0: Mean child/adolescent obesity prevalence did not change between the earliest and latest available years.

H_1: Mean prevalence in the latest year is higher than in the earliest year.

Hypothesis 3.2 Investigate effect of demographics on Childhood obesity

H_0: No difference in mean prevalence across key subgroups

H_1: There is at least one meaningful difference in subgroup disparity

- Statistical tests:

  One-way ANOVA per domain (sex, age, race/ethnicity, income/poverty).

- Graphs:
  - Total Trend- Line chart of overall child obesity prevalence
  - Sex Trend
  - Age Trend
  - Income/Poverty Trend

III. Data Sources and Preparation

3.1 BRFSS Adult Obesity Dataset (CDC, 2011–2023)

The adult obesity data were obtained from the Behavioral Risk Factor Surveillance System (BRFSS) Nutrition, Physical Activity, and Obesity dataset maintained by the Centers for Disease Control and Prevention (CDC). The dataset includes state-level observations from 2011 to 2023 covering all 50 states along with 3 U.S.

Territories. The cleaned dataset we used contains state-level estimates from 2011 through 2023. Each record includes the survey year, state abbreviation and name, the surveillance system, measure classification, the specific indicator question, the numeric estimate, confidence limits, sample size and stratification variables such as age groups (categorical), sex, education, income, and race/ethnicity.

The BRFSS data are based on self-reported height and weight collected via telephone survey; while this may underestimate true obesity prevalence, the survey's large sample size and uniform methodology across years make it well suited for assessing national trends and regional differences over time.

3.2 Sleep Health and Lifestyle Dataset (Kaggle Data for Machine Learning)

To explore how lifestyle and behavioral factors relate to BMI categories using predictive modeling, we used the Sleep Health and Lifestyle dataset from Kaggle. This dataset contains 374 individual-level observations, each corresponding to a unique person (Person.ID). Variables include demographic and occupational information (Gender, Age, Occupation), sleep and lifestyle factors, cardiometabolic measures (Systolic, Diastolic, Heartrate), a categorical BMI outcome (BMI.Category with levels Normal, Overweight, and Obese), and a sleep disorder field.

For our analysis, BMI.Category was treated as the target variable, and the remaining columns (excluding the person identifier) were used as predictors. The dataset did not contain missing values in the variables we used, so no imputation was required. Basic preprocessing steps included converting categorical variables (such as gender, occupation, and BMI category) into factors and ensuring that numeric variables (such as sleep duration, blood pressure, and daily steps) were in a suitable numeric format for modeling. Because this Kaggle dataset is synthetic and created primarily for educational purposes, the machine learning results are interpreted as an illustration of how lifestyle features relate to BMI categories in a modeling context.

3.3 Childhood Obesity Dataset (CDC/NCHS NHANES, 1988–2018, via Kaggle)

Childhood obesity analyses were conducted using a dataset distributed on Kaggle that reproduces the CDC/National Center for Health Statistics (NCHS) table "Obesity among children and adolescents aged 2–19 years, by selected characteristics: United States." The dataset file contains 840 rows and 16 variables,

summarizing pooled NHANES survey cycles from 1988–1994 through 2015–2018. Each row represents a combination of age panel, subgroup dimension, subgroup label, and survey period, along with the corresponding obesity prevalence estimate and its standard error. Because these values are derived from NHANES measuring height and weight, they reflect nationally representative, survey-weighted estimates.

For this project, we focused on the 2–19 years age panel and examined four subgroup dimensions relevant to our analyses: overall prevalence, sex, race and family income-to-poverty ratio. Because NHANES reports obesity estimates in pooled multi-year cycles, we analyzed these cycles in chronological order to assess changes over time and differences by sex, race/ethnicity and income. Unlike the synthetic lifestyle dataset used for machine learning, this NHANES-derived dataset reflects real-world, nationally representative measurements and allows for a robust comparison of childhood obesity trends and disparities across demographic and socioeconomic groups.

Methods

This project combines descriptive statistics, data visualization and basic predictive modeling across three complementary datasets. For adults, we used BRFSS state-level estimates of the "Percent of adults aged 18 years and older who have obesity" from 2011–2023 to describe national trends and regional differences. For children, we analyzed CDC-published NHANES estimates for ages 2–19 years across pooled survey cycles from 1988–1994 to 2015–2018 to examine how childhood obesity and related disparities have evolved over time. In between, we used an individual level lifestyle dataset from Kaggle to explore how sleep, stress, and activity-related variables relate to BMI category using simple machine-learning models.

For the BRFSS data, we first filtered the dataset to retain only rows where the indicator corresponded to adult obesity, and the stratification represented the total adult population rather than specific subgroups. We then organized the data by year, state and Census region and calculated summary statistics such as mean obesity prevalence by region and by year. Line graphs were used to visualize trends over time within each region, and a choropleth map was created for a recent year to highlight geographic variation across states. To support the

visual patterns statistically, we ran simple one-way and two way ANOVA tests to assess whether mean obesity prevalence differed significantly across Census regions and across age groups, and time using the BRFSS age-adjusted percentages as inputs.

For the Sleep Health and Lifestyle dataset, we treated BMI.Category (Normal, Overweight, Obese) as the outcome and used demographic, sleep, and lifestyle variables (e.g., age, gender, occupation, sleep duration, physical activity level, stress level, daily steps, blood pressure, heart rate) as predictors. The dataset contained no missing values in the fields we used, so no imputation was required. We randomly split the records into a training portion and a test portion, trained a baseline logistic regression model, multi linear regression model, a simple neural network, and a random forest classifier and then compared their accuracy on the test data. We also examined the random forest's variable importance to identify which factors were most strongly associated with BMI category in this simulated setting focusing particularly on sleep duration, stress, and physical activity.

For the childhood obesity analysis, we used the CDC/ NHANES summary table restricted to children and adolescents aged 2–19 years. Within this panel, we created analytic subsets for overall prevalence (total), sex, race and family income-to-poverty ratio. The NHANES estimates are reported for pooled multi-year survey cycles (e.g., 1988–1994, 1999–2002, 2003–2006), so we ordered these cycles chronologically and used them as the time axis for trend analyses. For each subgroup dimension, we summarized obesity prevalence across cycles and produced line or bar charts: an overall trend line for 2–19-year-olds, race/ethnicity trend lines to examine whether disparities have widened, an income-to-poverty gradient to assess socioeconomic differences, and age-group trends for 2–5, 6–11, and 12–19 years. For cross-sectional comparisons in the most recent period, we used ANOVA to test whether mean obesity prevalence differed significantly by race/ethnicity and by income group using the CDC published estimates as inputs.

IV. Results

U.S Obesity Temporal and Regional Trends 2011 - 2023.

For understanding regional and temporal trends in obesity, we analyze the dataset made from conducting surveys of adults aged above 18 by CDC Behavioral Risk Factor Surveillance System (BRFSS, 2011 - 2023). The obesity dataset (n = 4,134) post data cleaning includes records from 53 U.S. locations spanning 2011 to 2023 and captures six adult age groups (18–24, 25–34, 35–44, 45–54, 55–64 and 65 or older). Obesity prevalence varied widely, with a mean of 30.6%, median of 31.4% and standard deviation of 7.84, ranging from 3.8% to 59.7% and an interquartile range of 9.8. The distribution has moderate spread without extreme outliers as explained by skew and kurtosis. The dataset's 4,134 observations are evenly distributed across the 14 years from 2011 to 2023 (≈318 records per year) and similarly uniform across all six age groups (≈689 records per group).

Descriptive Statistics for Obesity Prevalence

N	Mean	Median	SD	Min	Max	Range	MAD	Skewness	Kurtosis	SE
4,134	30.59	31.40	7.84	3.8	59.7	55.9	7.41	-0.33	-0.11	0.12

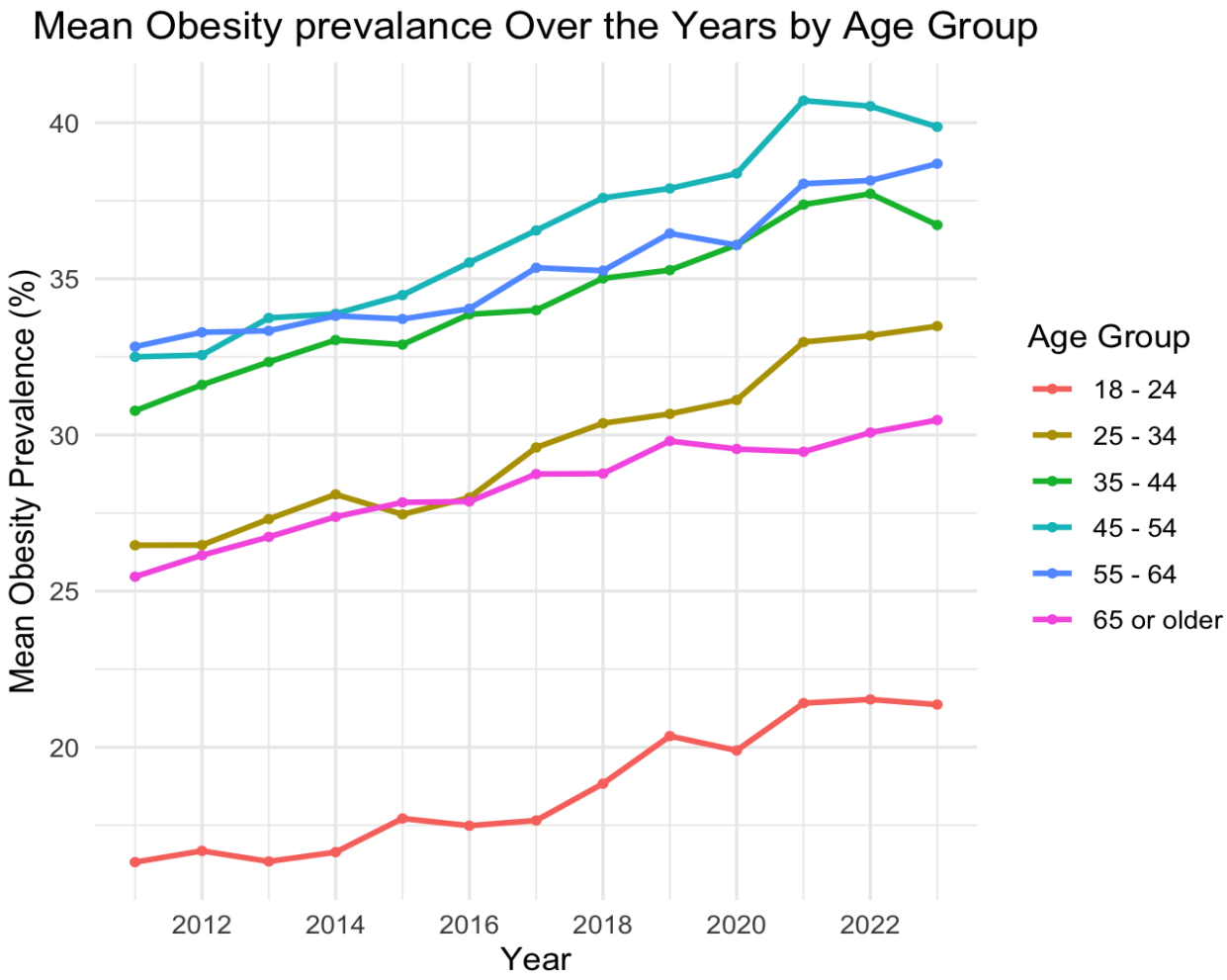
Mean Obesity prevalance Over the Years by Age Group
Mean Obesity Prevalence (%)
40
35
30
25
20
2012
2014
2016
2018
2020
2022
Year
Age Group
18 - 24
25 - 34
35 - 44
45 - 54
55 - 64
65 or older

Figure 1

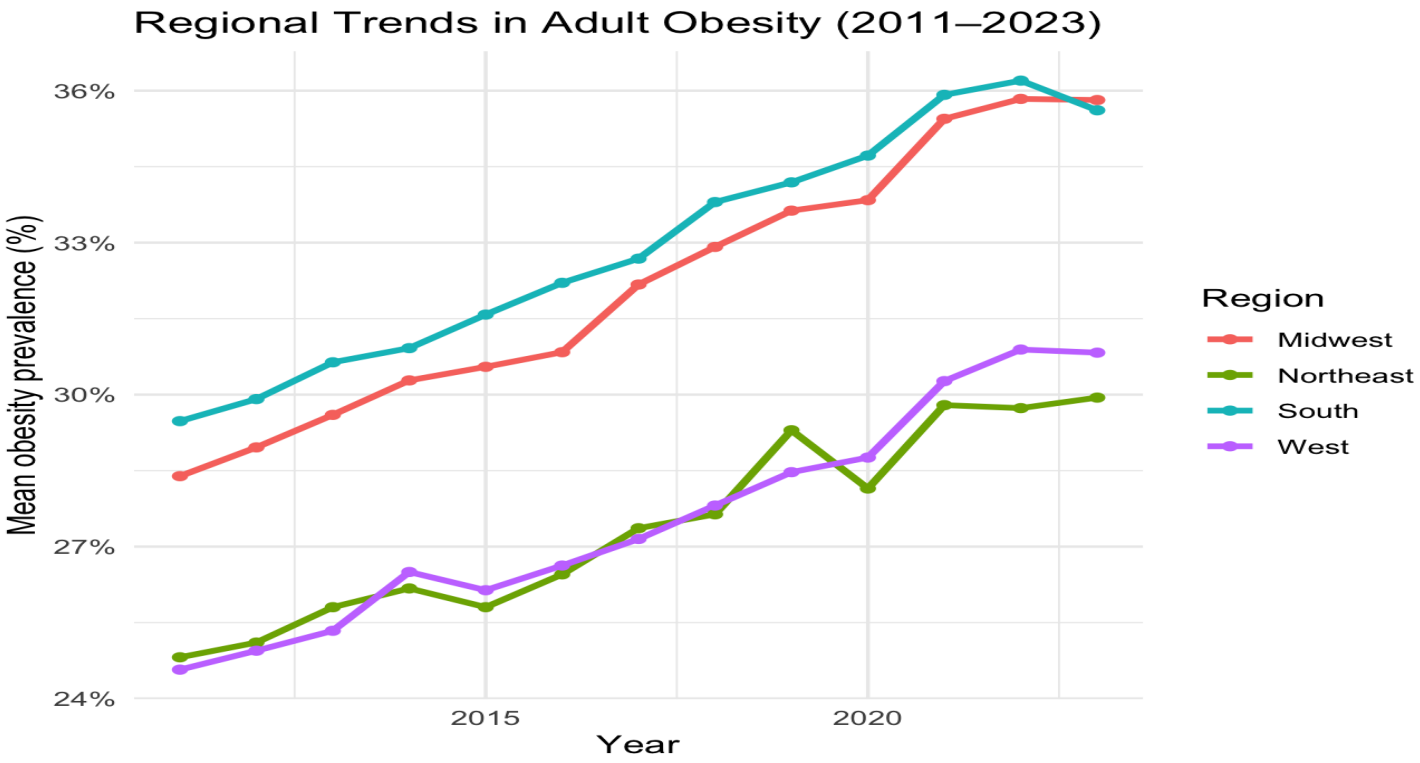

Figure 2

Looking at Figure 1 and 2, you can observe that the obesity prevalence has increased over time from 2011 to 2023 across all regions in the United States. In this analysis, we understand the key risks of obesity and if there is strong statistical evidence to suggest the same.

After running Anova tests obesity prevalence across various predictors in the data set, we have found:

Dominance of Age: Age is the most dominant factor of obesity. The factor (age) row has the largest Sum of Squares (426.2), suggesting that the differences between the various age groups account for the largest proportion of the total variation in the value variable. The year also has a substantial effect: The factor (year) row has the second-largest Sum of Squares (109.6), indicating a noticeable difference in the overall mean value between 2011 and 2023. The Interaction between the factors is minimal, the factor(year): factor(age) row has the smallest Sum of Squares (2.4), which suggests that the effect of age on value is very similar in 2011 and 2023. There is very little evidence of a significant interaction effect.

Two-way Anova Test	Degrees of Freedom	Sum of Squares	Mean Square
factor(year)	1	109.6	109.58
factor(age)	5	426.2	85.24
factor(year): factor(age)	5	2.4	0.48

Regional Trends (Hypothesis 1.2): Regional Trends also indicate a lower obesity prevalence in the west and the Northeast compared to the Midwest and South. The region explains the most variation in obesity prevalence (Sum Sq = 337.7). compared to a year which also explains a large amount of variation (Sum Sq = 238.2). The interaction is minimal (Sum Sq = 6.6), suggesting that the regional differences have been relatively constant over the years. The entire analysis is done on age adjusted data, that is the distribution of ages across regions and over the years has been kept constant to avoid any bias and analyze the impact of region and temporal differences without any bias from the most significant factor (age). Our findings closely mirror national CDC surveillance data. According to the 2023 CDC Adult Obesity Prevalence Maps, every U.S. state and territory now has an adult obesity prevalence above 20%. The Midwest (36.0%) and South (34.7%) have the highest obesity rates, followed by the West (29.1%) and Northeast (28.6%), an exact reflection of the patterns in Figure 2 (Centers for Disease Control and Prevention, n.d.).

Two - Way Anova Test	Degrees of Freedom	Sum of Squares	Mean Square
factor(region)	3	337.7	112.55
factor(year)	12	238.2	19.85
factor(region): factor(year)	36	6.6	0.18

1 Way Anova Tests	Degrees of Freedom	Sum of Squares	Mean Square	F-Value	Pr(>F)
factor(year)	12	18943	1578.6	27.49	<2e-16
Residuals	4277	245594	57.4		
factor(age)	5	157386	31477	1258	<2e-16
Residuals	4284	107152	25		
factor(region)	3	337.7	112.5	22.07	4e^-09
Residuals	48	244.8	5.1		

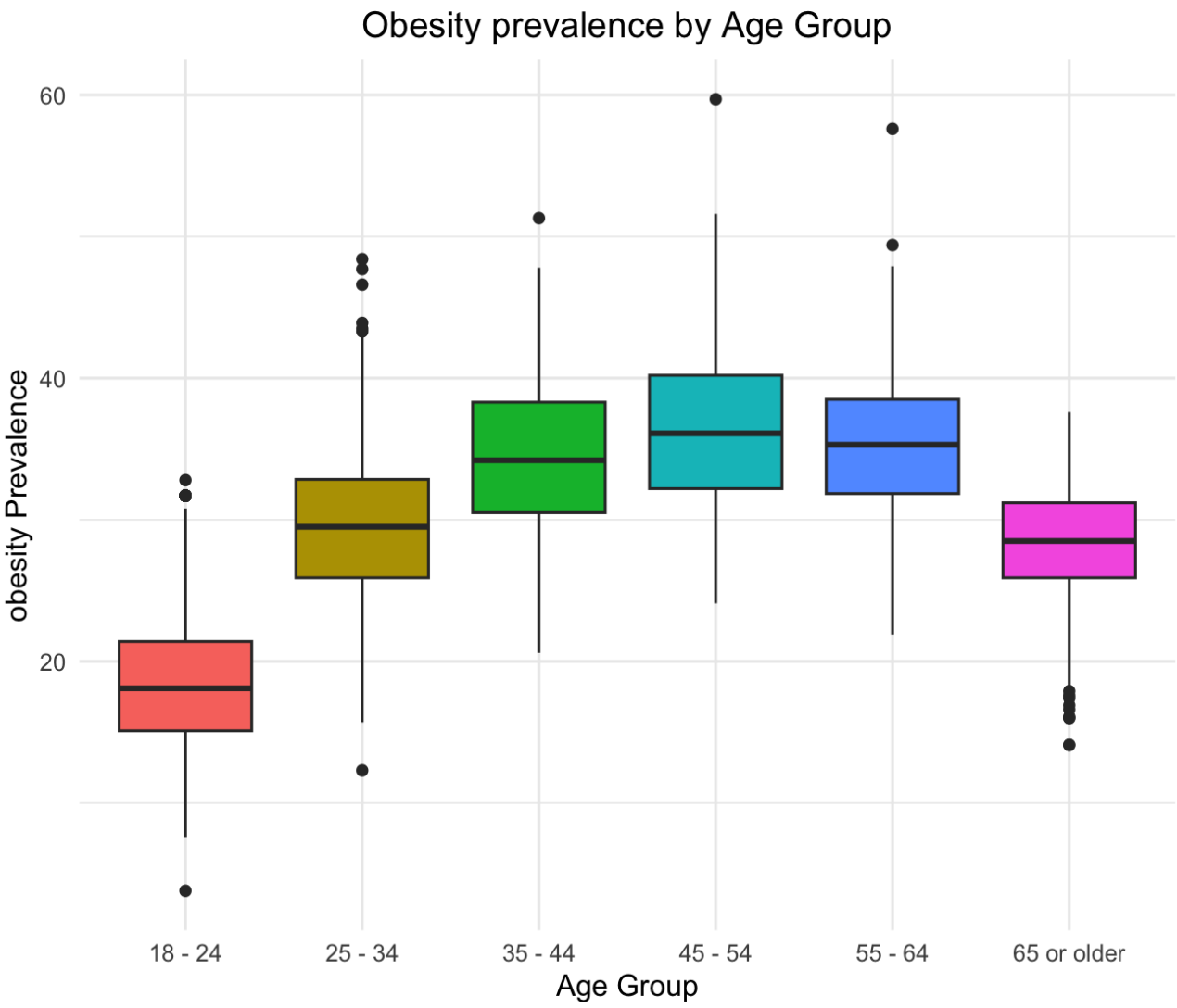

Figure 3

Peak Prevalence (Hypothesis 1.3): Figure 3 is a box plot of obesity prevalence based on age groups. From the box plot, obesity is at the highest for the age group '45-54', followed by '55-64' and '35-44'. Younger adults (18–24) consistently exhibit the lowest prevalence. This supports Hypothesis 1.3 regarding peak prevalence in mid-adulthood.

Summary of ANOVA tests:

The one-way ANOVA results showed significant differences in obesity levels across year, age, and region; however, the strength of these effects varied greatly between factors. Age demonstrated by far the largest impact ($F = 1258$, $p < 2e\text{-}16$), indicating substantial variation in obesity across age groups. Year showed a much smaller but still meaningful effect ($F = 27.49$, $p < 2e\text{-}16$), suggesting moderate changes in obesity over time. The region had the weakest of the three significant effects ($F = 22.07$, $p \approx 4e\text{-}09$), reflecting relatively smaller differences across geographic areas compared with age and year. Overall, while all three factors influence obesity, age is the dominant contributor, followed by year, with region having the least magnitude of effect.

The two-way ANOVA results support the same conclusions as the one-way tests: year, age, and region each have a significant effect on obesity. However, the interaction terms were not significant, meaning the changes over time were similar across both regions and age groups. In other words, although the overall levels differ, the pattern of change does not depend on where people live or their age category, confirming that the main effects are consistent and do not influence one another.

Investigate the Impact of Lifestyle Factors on Obesity - Machine Learning Model Performance Analysis (Objective 2)

For understanding the trends of obesity based on lifestyle like physical health activity, sleep and stress, multiple machine learning models were developed and pitted against each other on a dataset from Kaggle which contained several key lifestyle factors like, occupation, stress, sleep quality, hours of sleep, age, gender, and physical activity and obesity classes. The idea of the comparison is to understand if some machine learning models are significantly better than others in predicting obesity class. The dataset contains a predominantly normal-weight population (BMI range -18.5–24.9), with 57.8% (n = 216) classified as normal, followed by 39.6% (n = 148) in the overweight category (BMI range - 25.0–29.9)**,** and only 2.7% (n = 10) meeting criteria for obesity (BMI range ≥ 30.0).

Summary of Key Descriptive Statistics (N = 374)	
Measure	**Value**
Total sample size	**374**
Mean age (years)	**42.2**
Median age (years)	**43**
Age SD	**8.67**
Age range	**27–59**
Age IQR	**14.8**
Male (%)	**50.5% (n = 189)**
Female (%)	**49.5% (n = 185)**
Normal BMI (%)	**57.8% (n = 216)**
Overweight BMI (%)	**39.6% (n = 148)**
Obese BMI (%)	**2.7% (n = 10)**

Hypothesis 2.3 (Model Superiority): The results clearly demonstrate that the Random Forest models can predict a person's BMI category with a very high degree of accuracy. The most successful model was the Random Forest that used 5-fold cross-validation, which reached an accuracy of 98.67%. This validates the hypothesis that more complex machine learning models can effectively identify individuals at risk.

Hypothesis 2.2 (Feature Importance)

Having already established that age is a significant factor, we wanted to shift our focus to all the various rich lifestyle factors available in the dataset. After training the model on all the features, the importance () function was used to extract the raw importance scores for each predictor. The feature of importance for the Random Forest models was evaluated using two standard metrics: Mean Decrease in Gini and Mean Decrease in Accuracy. The results were then visualized using the varImpPlot () function, which generates a dot chart that ranks the features by their importance, allowing for a clear, visual comparison of their predictive power.

The analysis provided strong evidence that **occupation is the most influential predictor of obesity** among all the variables considered based on Mean Decrease in Gini and Mean Decrease Accuracy. The same analysis was performed using critical lifestyle predictors like sleep duration, physical activity level, and stress

level. Sleep duration had the greatest impact on Obesity Class. These factors were chosen specifically because these could be critical habits that can be incorporated into daily routine to reduce the risk of obesity.

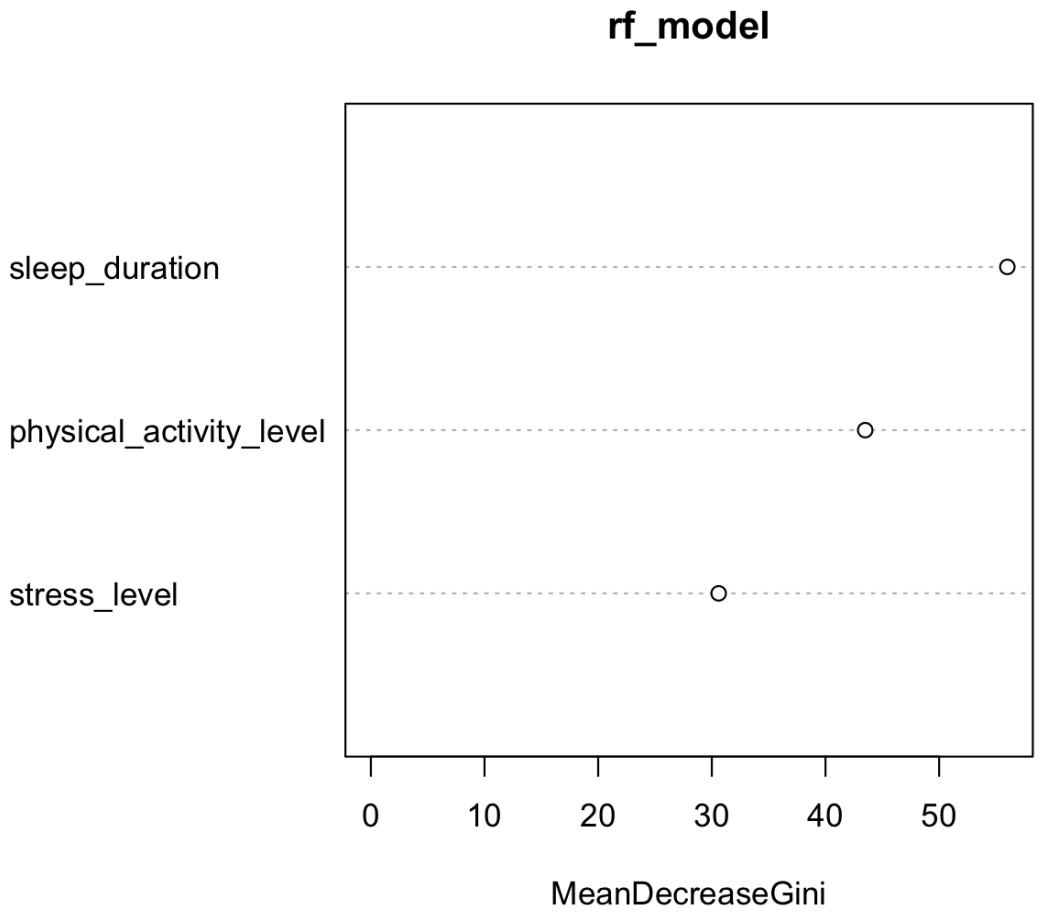

Figure 4

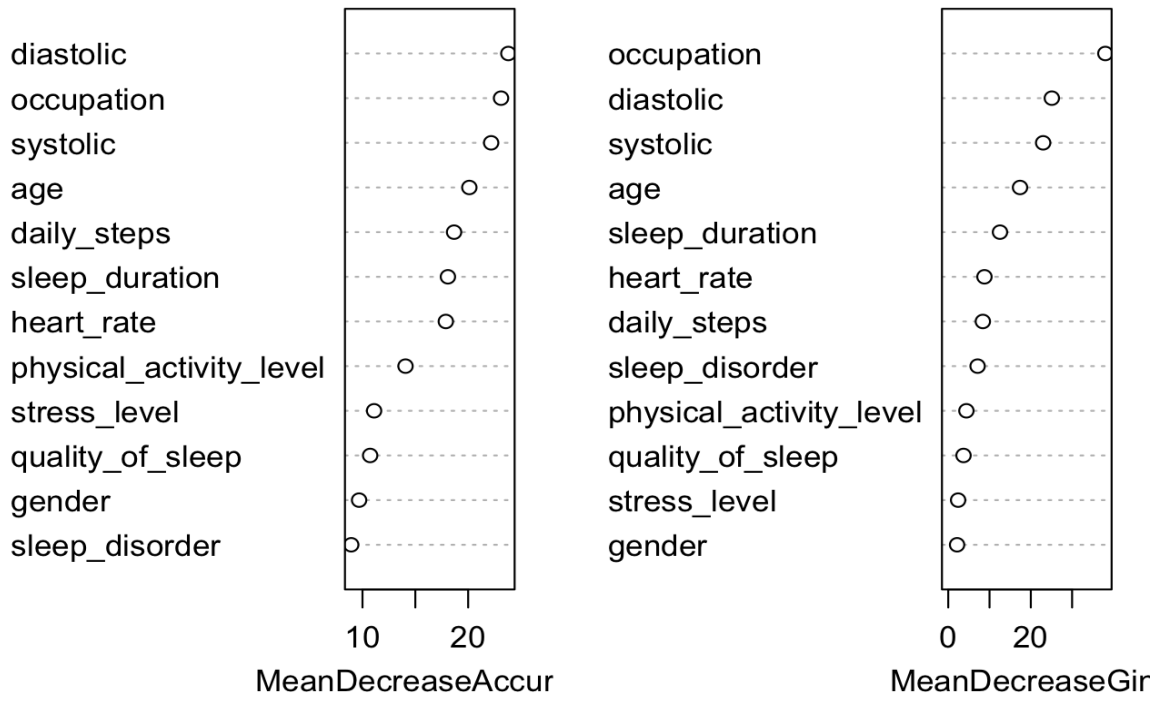

Figure 5

1. Variable Importance Plot: The variable importance plot generated from Random Forest model that included all features, which clearly ranked 'Occupation' at the very top. This indicates that occupation has the highest score in 'Mean Decrease Gini,' a metric that measures how much a variable contributes to the purity of the decision tree nodes in the Random Forest. A higher value means the model relies more heavily on that variable to make accurate predictions. Therefore, occupation is the single most important factor in predicting an individual's BMI category in this model.

2. Chi-Squared Test: To statistically validate the importance of occupation, a Chi-squared test was conducted. The test resulted in a p-value of less than 2.2e-16, which is extremely statistically significant. This result allows us to reject the null hypothesis that there is no association between occupation and BMI. It

confirms that the relationship between a person's job and their obesity status is not due to random chances in our dataset.

Model	Feature Set	Accuracy
Neural Network (8, 4)	Sleep Duration, Stress Level, Physical Activity Level.	69.9%
Logistic Regression	Sleep Duration, Stress Level, Physical Activity Level.	70.79%
Random Forest - Basic	Sleep Duration, Stress Level, Physical Activity Level.	89.33%
Random Forest (with 5-fold Cross-Validation)	Sleep Duration, Stress Level, Physical Activity Level.	98.67%

Table 1: Model Performance Comparison

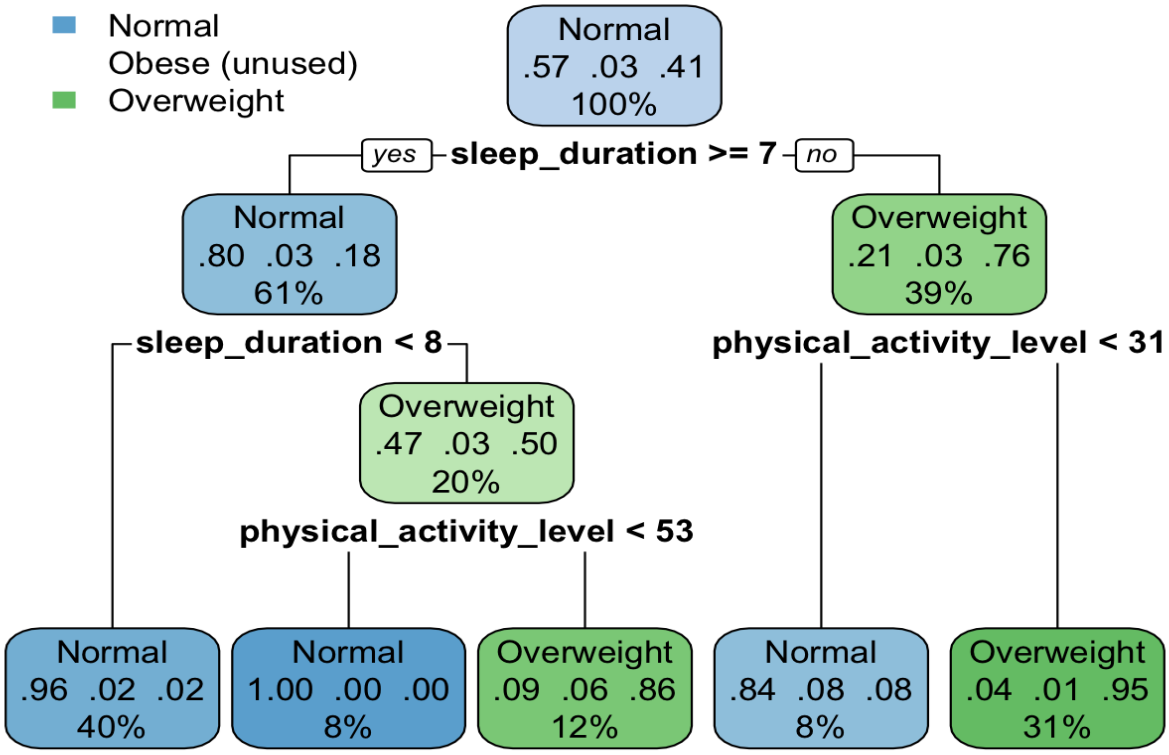

Figure 6

Due to the lack of publicly available test data, the dataset was split 70/30 and used for both training and testing model performance. Some key findings include:

Highly Accurate Prediction is Achieved: The results clearly demonstrate that the Random Forest models can predict a person's BMI category with a very high degree of accuracy. The most successful model was the Random Forest that used 5-fold cross-validation, which reached an accuracy of 98.67% on the test data (source). This result validates the hypothesis that more complex machine learning models can effectively identify individuals at risk of obesity based on their daily habits.

Comparison of Simple vs. Complex Models: The analysis provides a compelling look at the trade-offs between model complexity and performance.

a) **The Top Performer:** The cross-validated Random Forest significantly outperformed all other models, highlighting the power of ensemble methods and cross-validation for this type of data.

b) **The Baseline Complex Model:** The basic Random Forest, even without cross-validation, achieved a respectable 89.33% accuracy, demonstrating the inherent strength of this algorithm for the task.

c) The Weaker Performers: Both the Neural Network (69.9%) and Logistic Regression (70.79%) models underperformed in comparison. This is due to lack of strong enough signals, especially in the selected feature set (physical activity, stress levels and sleep duration). The comparatively poor performance of our **neural** network (70% accuracy) does not imply neural nets are inherently linear or limited. It's a reflection on our dataset and the features we have selected to train and tune the neural network. In practice, neural networks can achieve very high accuracy on similar classification tasks when properly configured. For example, Helforoush and Sayyad (2024) developed a hybrid artificial neural network optimized with Particle Swarm Optimization (ANN-PSO) to predict obesity risk, achieving 92% accuracy outperforming traditional regression methods. On increasing the feature set to all features, we have identified and verified the same, both neural network and logistic regression performed significantly better.

Model	Feature Set	Accuracy

Neural Network (8, 4)	All Features	97.35%
Logistic Regression	All Features	97.33
Random Forest - Basic	All Features	96%

2. **Hypothesis 2.1 (Core Habit Importance):** When focusing only on the core lifestyle habits of sleep, stress, and physical activity, the analysis identifies **sleep duration as the most critical factor**. The Random Forest model's importance metrics show that 'Sleep Duration' is the most powerful predictor. It has the highest MeanDecreaseAccuracy' (55.25) and 'MeanDecreaseGini' (59.12) scores. This means that, within this context, knowing a person's sleep duration provides the most information to the model for predicting whether they are of a normal weight, overweight, or obese, more so than their stress or physical activity levels.

In conclusion, the analysis effectively demonstrates that a finely tuned complex model like a cross-validated Random Forest can achieve near-perfect accuracy, and that the choice of model has a significant impact on predictive performance.

The Social Determinants of Childhood Obesity (Objective 3)

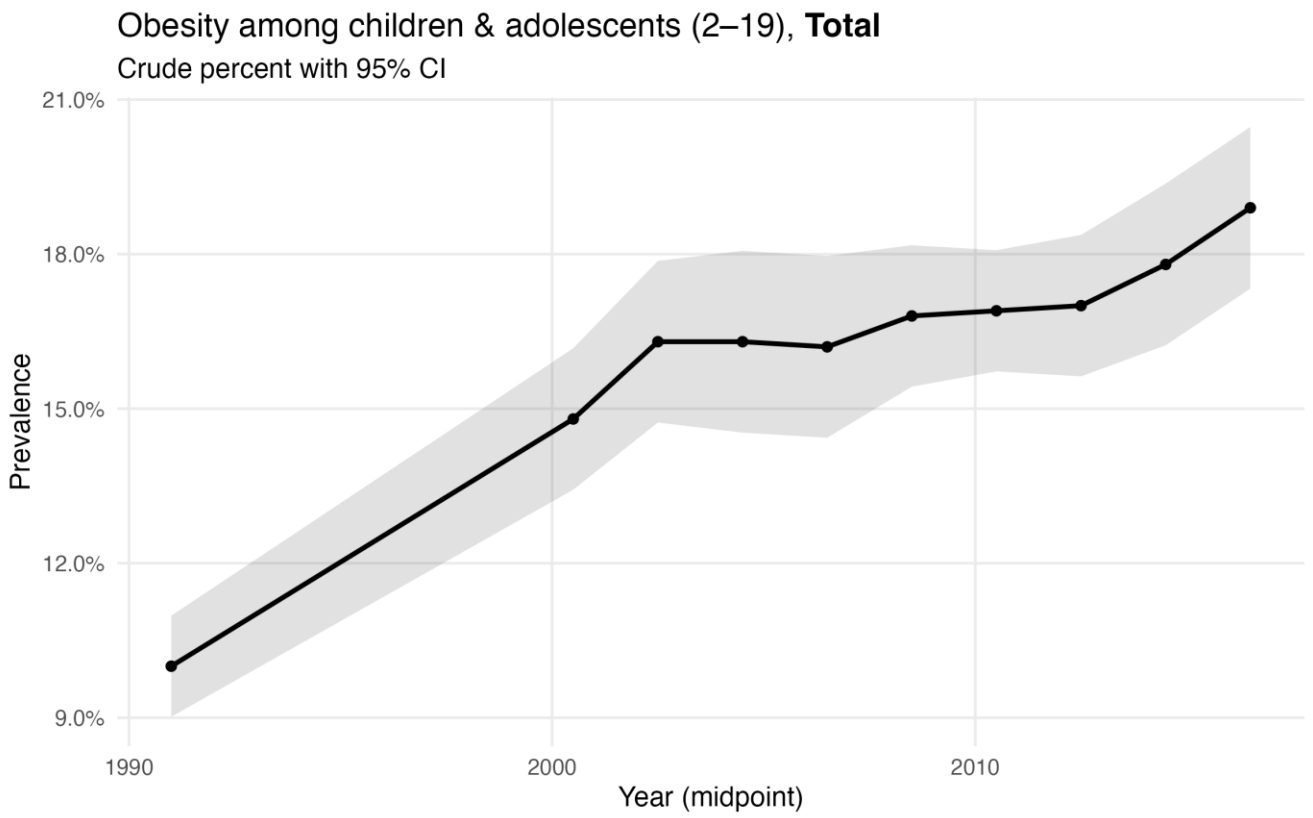

Figure 7

Hypothesis 3.1 (Temporal Trend)

The analysis of the data provides a stark and unambiguous result, illustrated in Figure 7. The prevalence of obesity among US children and adolescents aged 2-19 years increased from ~10% in 1988-1994 to ~19% in 2015-2018, nearly doubling over the observed period. The null hypothesis (H0) is rejected. This increase over three decades confirms the statement that childhood obesity has transformed from a minor health concern into a pervasive population-level epidemic.

Hypothesis 3.2 (Subgroup Disparities): The null hypothesis (H0) is rejected. The data visualized in Table 2 and 3 along with Figures 9, 10 and 11, confirm profound and statistically significant disparities across subgroups.

Subgroup	Characteristic	Prevalence (%)
Overall	**Ages 2-19**	18.5**%**
Age	2-5 years	13.9%
	6-11 years	18.4%
	12-19 years	20.6%
Race/Ethnicity	Hispanic	25.8%
	Non-Hispanic Black	22.0%
	Non-Hispanic White	14.1%
	Non-Hispanic Asian	11.0%

Table 2: Prevalence of Childhood Obesity (Ages 2-19) by Key Demographic and Socioeconomic Characteristics in 2015-2016

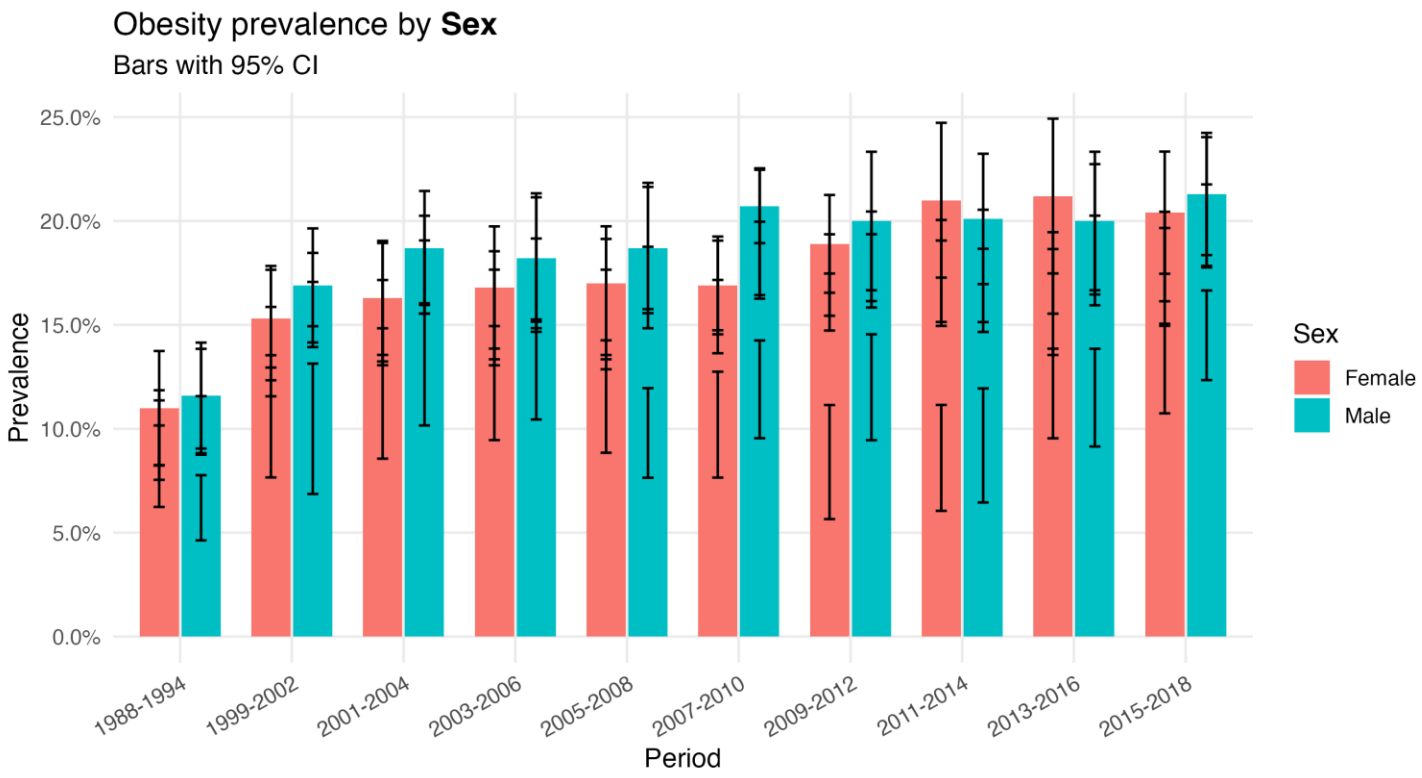

Figure 8

domain	method	stat	df	p_val
Sex	Weighted ANOVA	0.6012546	1	4.68e-01
Age	Q-test (IV-weighted)	27.6601307	2	9.86e-07
Race/Ethnicity	Weighted ANOVA	16.0912153	13	1.98e-11
Income/Poverty	Weighted ANOVA	11.6859030	3	7.13e-04

Table 3: Anova test for various socio-economic factors

Table 3, showcase that factors like age, race/ethnicity and income/poverty have statistically significant impact on childhood obesity (p value < 0.05). We have evaluated each of these factors in the following section.

The Sex Gradient: Not statically significant

As seen in Figure 8, there was no significant difference by Sex (p value = 0.468). The high p value of 0.468 indicates that unlike the other factors sex is not a statistically significant driver of obesity in childhood. The minor variations seen in the graph between boys and girls are attributable to random chance, not a true population-level difference. This implies that the powerful, systemic factors driving the epidemic, such as the socioeconomic and environmental conditions identified in this analysis, affect both boys and girls with equal force during these formative years.

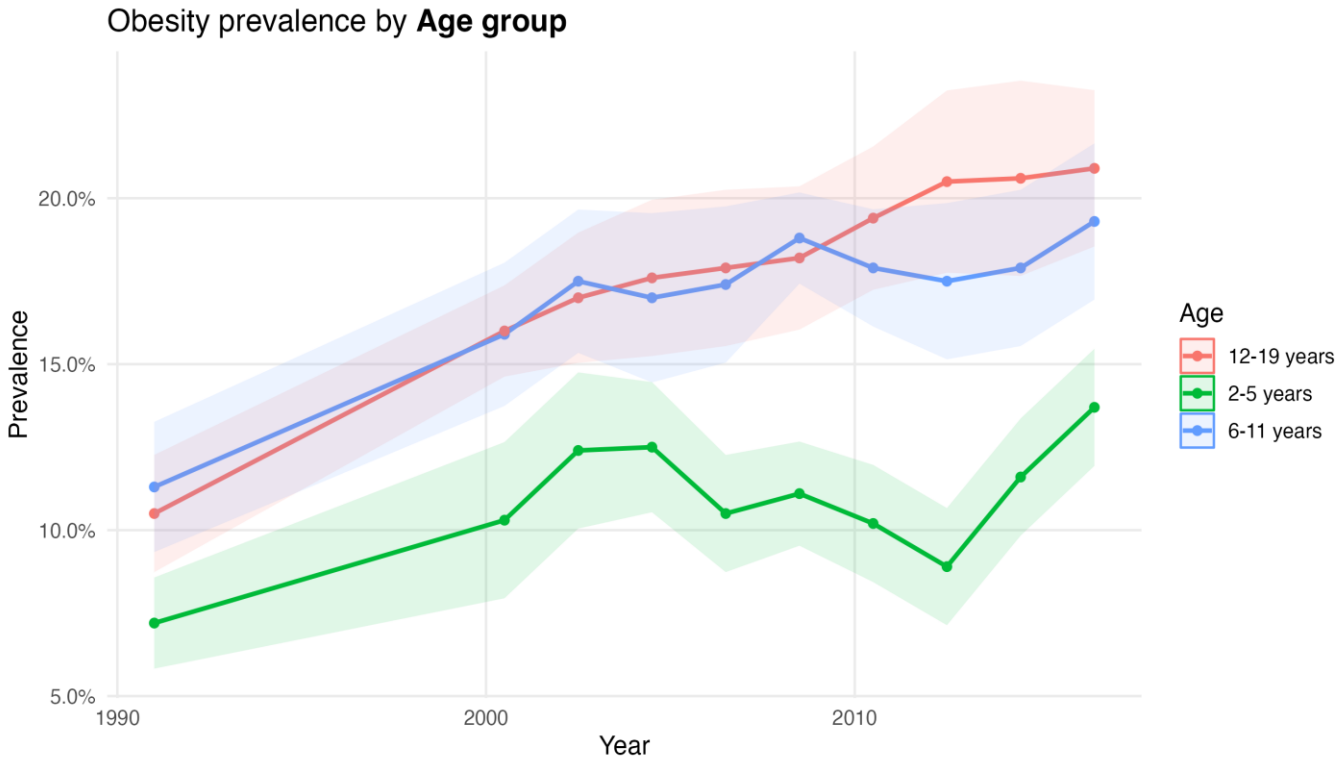

Figure 9

The Age Gradient: Risk Accelerates During School Age

The data shows a clear age gradient, with prevalence increasing as children get older: 13.9% (ages 2-5), 18.4% (ages 6-11) and 20.6% (ages 12-19) in years 2015–2016. The most significant jump, a 4.5% increase occurs between the 2-5 age group and the 6-11 age group. This transition maps directly to a critical life-course change: the move from a controlled home environment into the broader school system. In this new environment, children are exposed to a host of new factors, including the variable nutritional quality of school lunches, the availability of vending machines, the influence of peer-group food culture and reduced parental supervision over dietary choices. This finding strongly suggests that the school environment itself is a critical, high leverage setting for public health intervention.

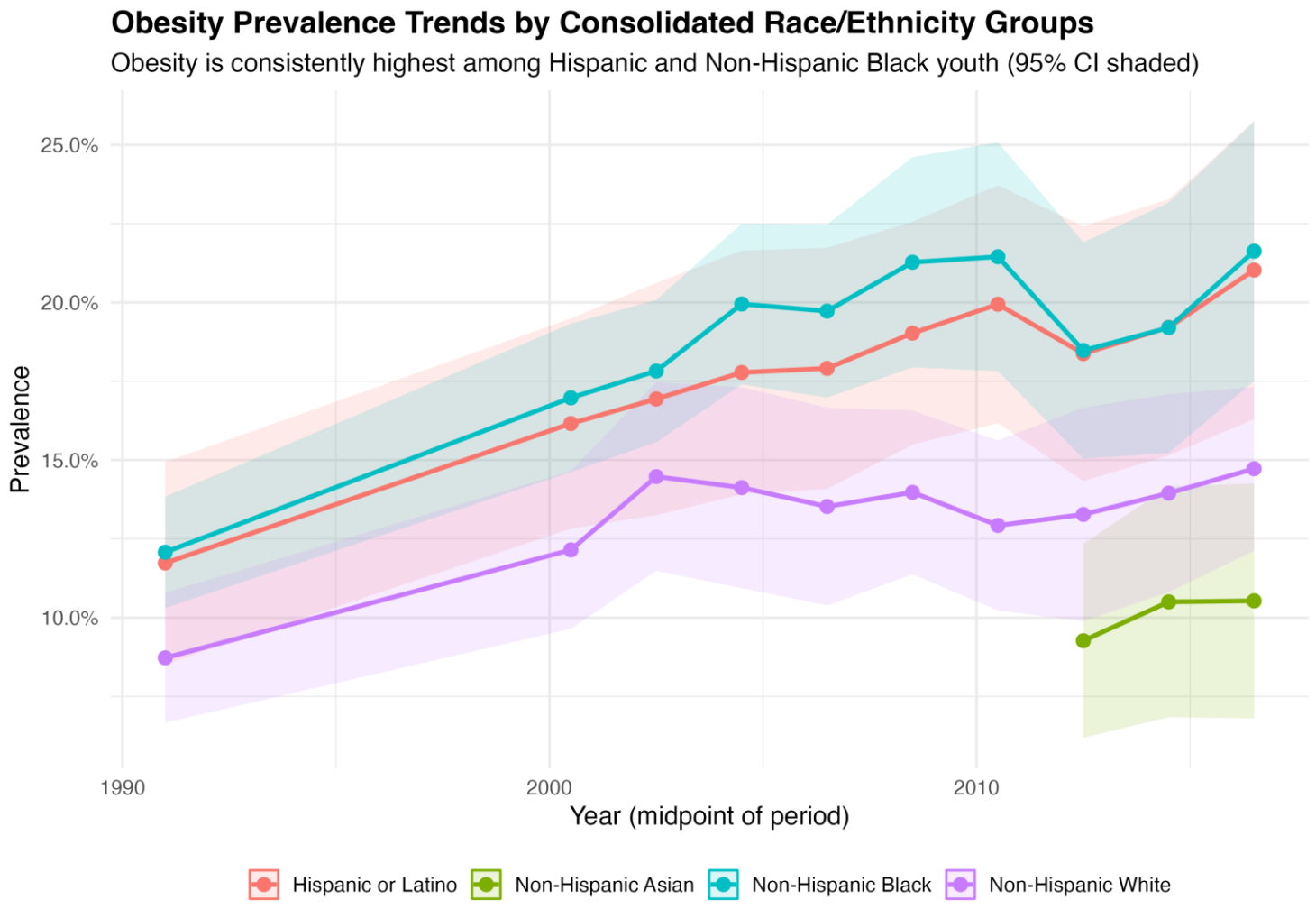

Figure 10

The Racial and Ethnic Gradient: Evidence of Structural Inequity

The disparities by race and ethnicity are profound and statistically significant, as confirmed by the p value = 1.98e-11. The figure 10 shows that there is a consistent and widening gap between Hispanic and non-Hispanic Black children and their White and Asian peers. The prevalence curves for Hispanic/Latino (Red line) and Non-Hispanic Black (Blue line) youth track closely together and remain significantly higher than all other groups over the entire period. By 2018, obesity rates for Black and Hispanic youth converged at approximately 21-22%, whereas Non-Hispanic White youth (Purple line) remained near 15%, and Non-Hispanic Asian youth (Green line) had the lowest prevalence at roughly 11%. These disparities are statistical proofs of structural inequities including unequal access to healthy food, recreational spaces, safe neighborhoods and economic stability.

The graph shows what is happening but the study by Aaron and Stanford (2021) explains the underlying causes. The significantly higher rates among Black and Hispanic youth align with Aaron and Stanford's findings that companies producing processed foods have historically engaged in disproportionate marketing towards Black communities. The study also highlights that processed nutritionally poor foods are often lower in cost which makes them more accessible to populations with lower socioeconomic status (GlobalData Healthcare, 2023).

The intersection of these findings suggests an unhealthy trajectory. As Aaron and Stanford note, unhealthy BMI status is often established at an early age and continues into adulthood. The high prevalence rates shown in the graph for Black and Hispanic children are not just childhood problems but they are the precursors to the high adult obesity rates identified in your Objective 1 analysis. Therefore, eliminating disparities requires more than just individual dietary advice. It requires regulatory interventions to curb targeted marketing and economic policies to make healthy food affordable for all communities.

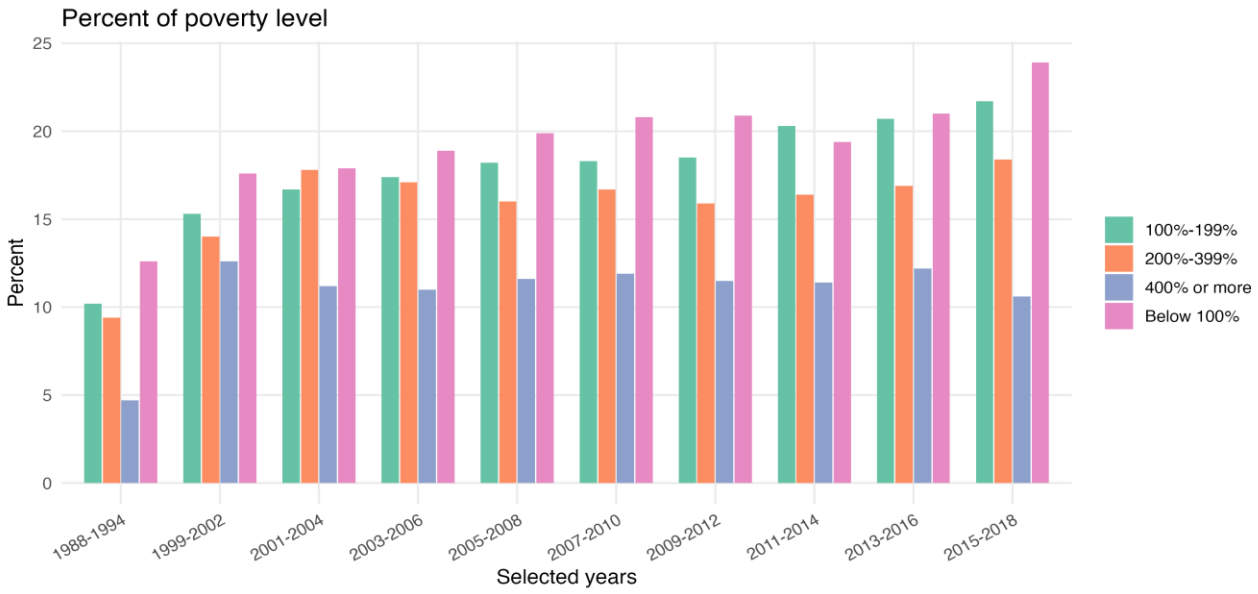

Figure 11

The Socioeconomic Gradient

The data shows that children in families living in poverty (<100% FPL) have the highest prevalence while children in high-income families (>400% FPL). The most powerful finding from this analysis is the stark, dose-response relationship between family income and obesity prevalence, illustrated clearly in Figure 11 and confirmed by the highly significant p-value of 7.13e-4. The data is unequivocal: children in families living in poverty have the highest prevalence, while children in high-income families have the lowest prevalence.

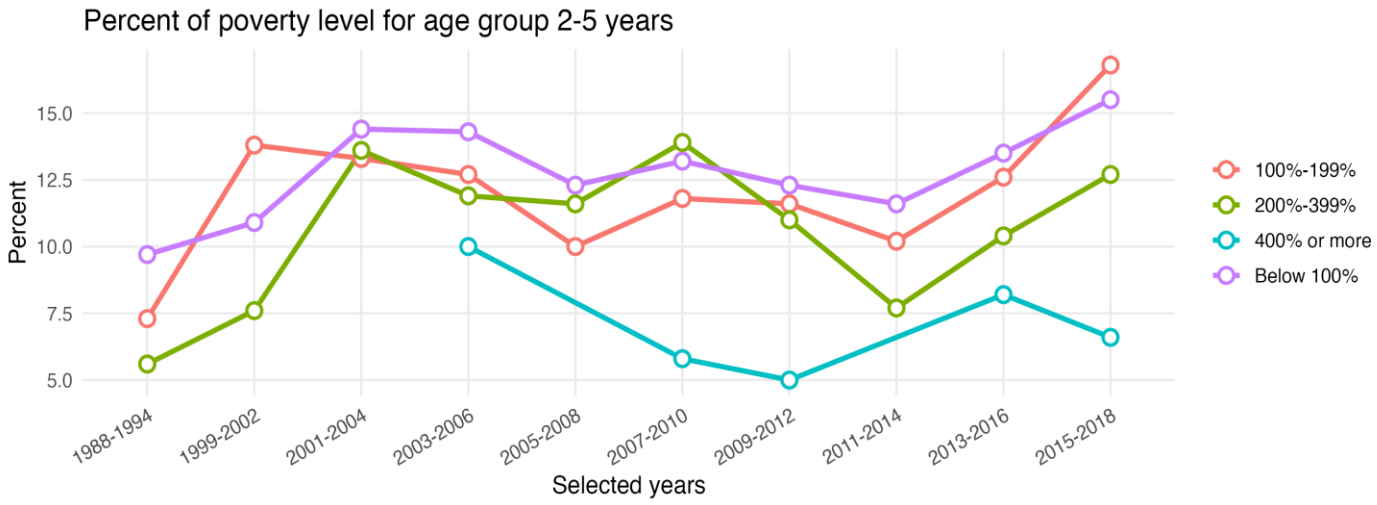

Figure 12

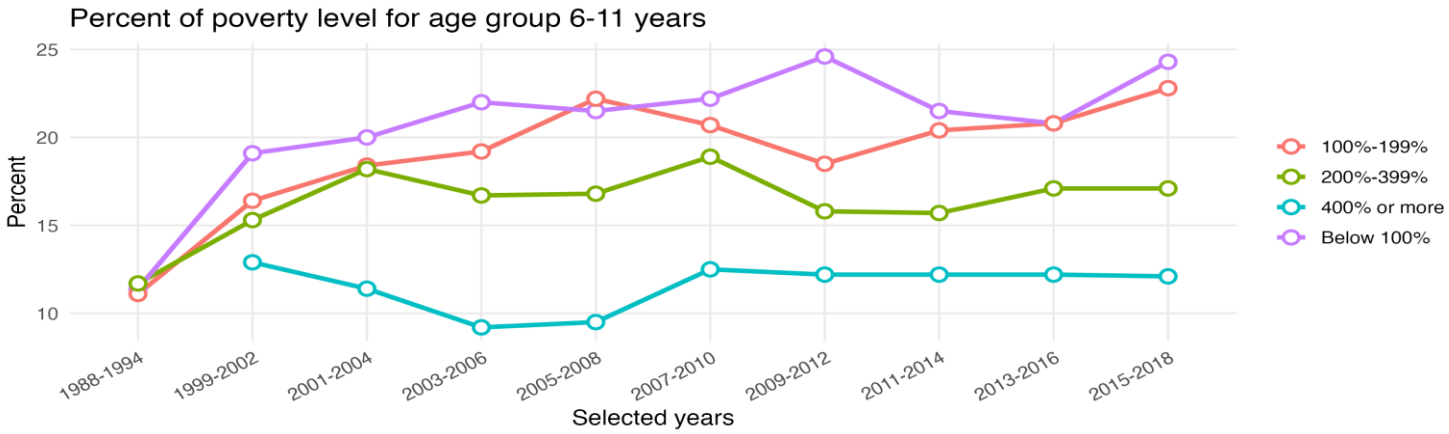

Figure 13

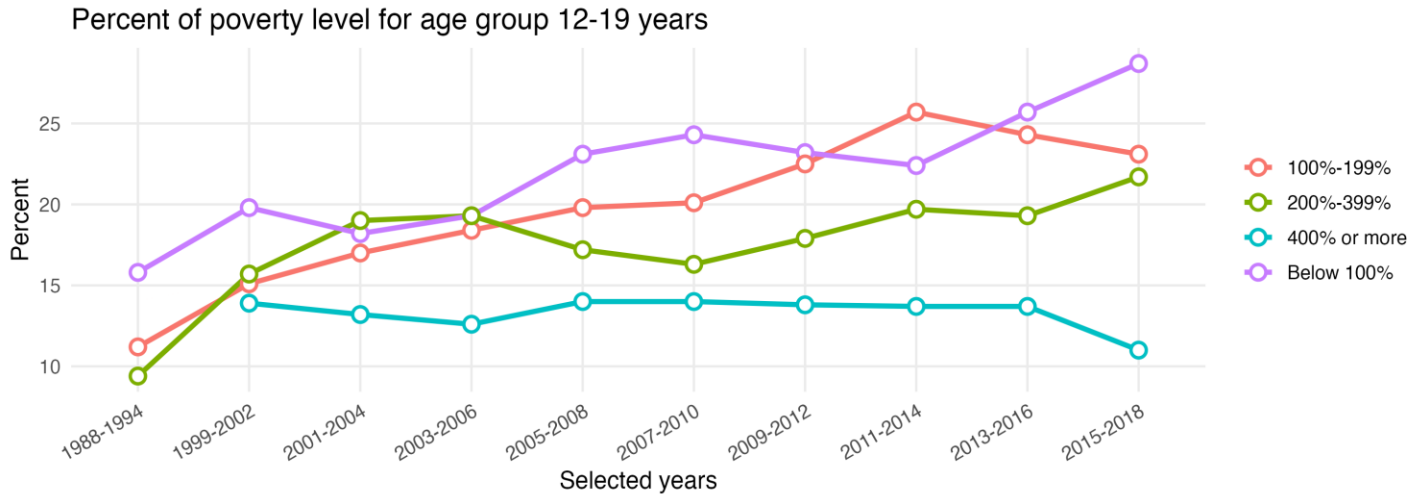

Figure 14

A deeper analysis of graphs plotting poverty by age groups reveals a crucial insight. In the 2-5 age group (Figure 12), the prevalence lines for different poverty levels are clustered together. However, in the 6-11 age group (Figure 13) the gap widens dramatically; the "Below 100%" poverty line separates and remains high while the "400% or more" (high income) line remains low. This visual evidence strongly suggests that the health gap between high-income and low-income children accelerates and solidifies during elementary and middle school years. Higher-income families can afford healthier and often more expensive foods and can live in safer neighborhoods with access to parks, playgrounds and grocery stores. They can pay for extracurricular

sports and activities. Lower-income families are systematically denied these choices. These findings confirm that childhood obesity is at its core **a disease of socioeconomic inequality.**

V. Discussion & Comparative Analysis: Two Epidemics, One Social Gradient

Taken together, the three objectives of this project reveal a unified picture of obesity as a life course epidemic driven by structural and behavioral factors that reinforce one another over time. By integrating national adult surveillance data (Objective 1), lifestyle-based machine learning models (Objective 2) and NHANES childhood obesity trends (Objective 3), a common gradient emerges that obesity risk is shaped early, strengthened during adolescence, and compounded across adulthood, especially among populations facing socioeconomic and racial inequities.

Prevalence and Age: The Life Course of Risk

A direct comparison of prevalence shows that adult obesity is roughly double that of youth obesity. This illustrates that obesity is a chronic condition that accumulates over time. Our project's data paints a clear picture of this accumulation: Objective 3 (Childhood) shows risk is acquired early and accelerates during the school-age years, rising from 13.9% (ages 2-5) to 20.6% (ages 12-19). Objective 1 (Adulthood) shows this risk of compounds over decades, peaking in middle age (45-54 age group).

In the BRFSS adult data, the South and Midwest show the highest age-adjusted obesity prevalence, while the Northeast and West remain lower, patterns that closely match CDC Adult Obesity Maps. In the

NHANES childhood data, a similar gradient appears along racial/ethnic and socioeconomic lines: Hispanic and non-Hispanic Black children have much higher obesity prevalence than non-Hispanic White, and there is a strong relationship with family income with the highest rates among children living below 100% of the federal poverty level and the lowest among those in households above 400% FPL. These disparities widen with age, especially during late childhood and adolescence, indicating that inequities in school environments, neighborhood conditions, and access to healthy foods and safe spaces compound over time.

VI. Limitations, Insights and Impact

Limitations

This section will transparently acknowledge any limitations inherent in the analysis. The conclusions drawn are contingent upon the datasets used.

Data Source Mismatches: The three objectives used three different datasets with different methodologies.

First, Objective 1 relies on BRFSS, which uses self-reported height and weight. Self-report tends to underestimate BMI compared with measured values, so the adult obesity prevalences we report are likely conservative. The temporal and regional patterns remain valid, but the absolute levels should be interpreted as minimum estimates of the true burden.

Second, Objective 2 used a relatively small “sleep and lifestyle” dataset for the analysis of lifestyle factors, which was not age-adjusted. As the analysis of the first dataset established age as a significant factor in obesity prevalence, the lack of age adjustment in the second dataset introduces a potential confounding variable. While the analytical methodologies employed are robust, the results are constrained by the nature of the available data. Future work using larger, representative datasets with age adjustment and external validation would be valuable to confirm these patterns.

Third, Objective 3 used NHANES-based childhood obesity estimates (via the CDC Kaggle file). While these data are measured rather than self-reported and are nationally representative, they are cross-sectional, pooling multiple survey cycles. We also focused on a limited set of social determinants (income and race/ethnicity); other important factors such as neighborhood food environments, built environment and family-level stressors were not directly modeled.

Insights and Impact: A Data-Driven Logic Model

Despite these limitations, the three objectives together provide a powerful, data-driven logic model for understanding and addressing obesity. The combined results can be used to create a highly targeted and data-driven public health campaign.

Regional and state-level BRFSS analyses show where adult obesity is most concentrated. Figures 15 and 16 illustrate how state-level bar charts and heatmaps can identify geographic "hot spots" such as West Virginia, Mississippi, and Arkansas, as well as consistently lower-prevalence states such as Colorado. Public health departments could use similar maps to prioritize resources, focusing on surveillance, funding, and intervention programs in the highest-burden states and regions. For example, public health departments of every state could use regional analysis and machine learning results to identify high-risk areas to understand and develop interventions focusing on lifestyle drivers that reduce the risk of obesity. This targeted intervention ensures that public funds are invested in interventions focusing on sleep, stress levels, and physical activity factors for each sub-population.

Imagine a scenario where a public health or federal organization is tasked to tackle obesity. Instead of focusing on a uniform approach for all the populations, they can come up with targeted interventions that suit each section of the population by using framework as follows:

Step 1 – Identify where to act: Use BRFSS regional and state-level maps (as in Figures 13 and 14) to locate high-burden states and communities.

Step 2 – Identify what to target: Apply lifestyle focused models to understand which behaviors like sleep, physical activity, and stress are most strongly related to obesity in those specific populations.

Step 3 – Identify whom to prioritize: Use child and adolescent data to focus on low-income and racially marginalized groups, particularly during the elementary and middle-school years when disparities widen.

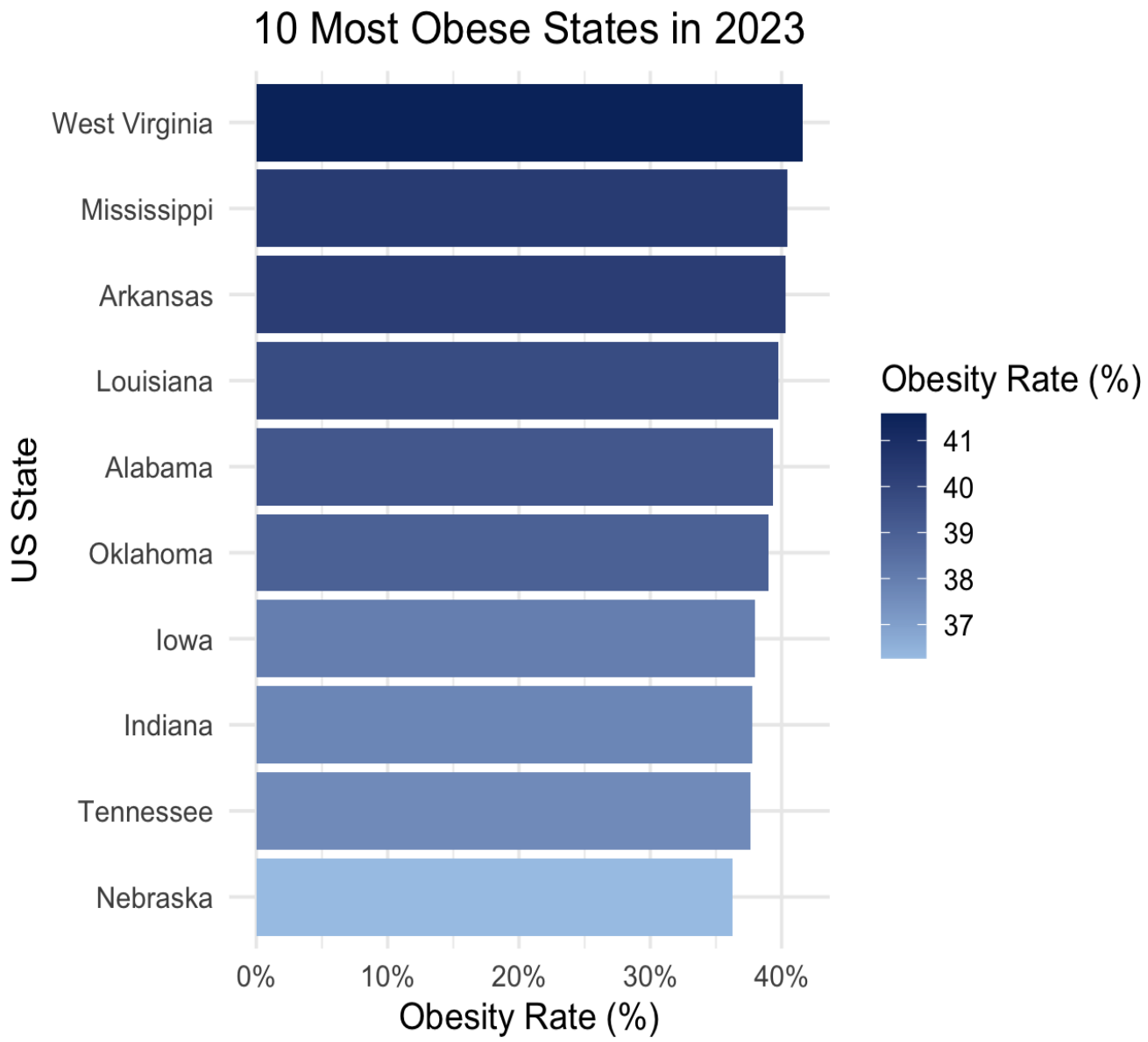

Figure 15

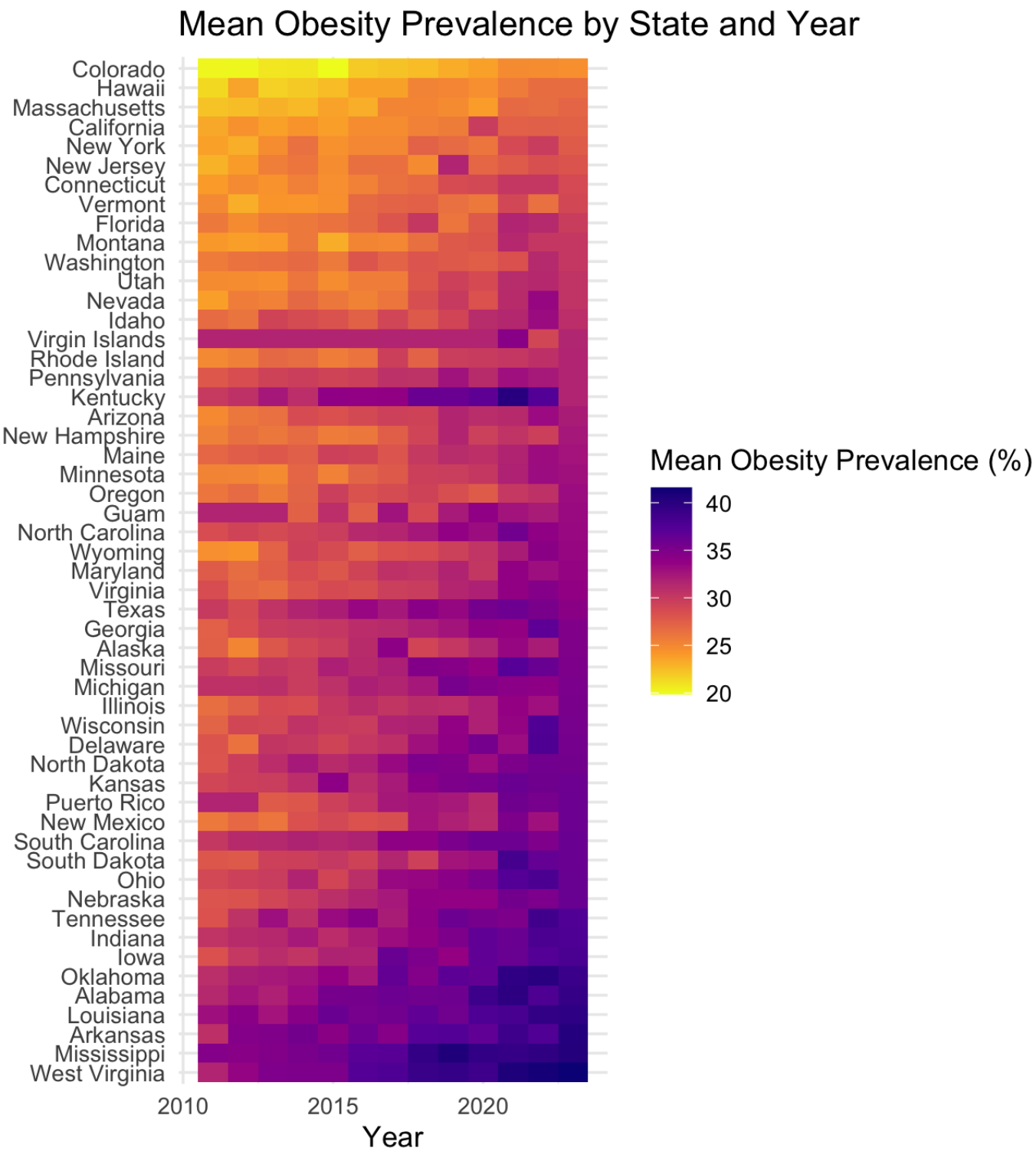
Mean Obesity Prevalence by State and Year
Colorado
Hawaii
Massachusetts
California
New York
New Jersey
Connecticut
Vermont
Florida
Montana
Washington
Utah
Nevada
Idaho
Virgin Islands
Rhode Island
Pennsylvania
Kentucky
Arizona
New Hampshire
Maine
Minnesota
Oregon
Guam
North Carolina
Wyoming
Maryland
Virginia
Texas
Georgia
Alaska
Missouri
Michigan
Illinois
Wisconsin
Delaware
North Dakota
Kansas
Puerto Rico
New Mexico
South Carolina
South Dakota
Ohio
Nebraska
Tennessee
Indiana
Iowa
Oklahoma
Alabama
Louisiana
Arkansas
Mississippi
West Virginia
2010
2015
2020
Year
Mean Obesity Prevalence (%)
40
35
30
25
20

Figure 16

State-level findings also align with Figures 15 and 16. In 2023, Colorado remained among the lowest-obesity states (<25%), while West Virginia, Mississippi, and Arkansas exhibited prevalence levels of 40% or greater, matching our top-state bar chart and confirming persistent high-obesity clusters in the South regions. CDC distributions showing most states falling between 30–40% also closely mirror the patterns seen in our heatmap.

VII. Conclusion

This multi-level analysis of obesity in the United States synthesizing three distinct analytical objectives, reveals a powerful and consistent narrative. The US obesity epidemic is not a uniform crisis but a disease of inequality, characterized by a steep and persistent social gradient. This analysis has demonstrated that this gradient is imprinted in early childhood, where socioeconomic status is the strongest predictor of obesity risk. The data shows this risk is not static; it accumulates across the life course accelerating during the school-age years and compounding into the high prevalence rates seen in middle-aged adults.

In addition, this trajectory is compounded by a host of other structural factors. We found that lifelong, systemic inequities create racial and ethnic disparities that begin in toddlerhood and are maintained into adulthood. We also found that for adults, structural determinants like occupation, a proxy for income, stress, and sedentary time are more predictive of obesity than many individual behaviors.

The inescapable conclusion is that public health strategies focused exclusively on individual behavior are aimed at the symptoms, not the disease. The findings of this report provide a clear, data-driven mandate to shift our focus. Effective and equitable solutions must therefore look beyond individual behavior to dismantle the systemic drivers of the epidemic: the economic, social, and environmental conditions that create and reinforce health inequities, placing a healthy life beyond the reach of millions.

References

Anekwe, C. V., Jarrell, A. R., Townsend, M. J., Gaudier, G. I., Hiserodt, J. M., & Stanford, F. C. (2020). Socioeconomics of Obesity. Current obesity reports, 9(3), 272–279. https://doi.org/10.1007/s13679-020-00398-7

Autret, K., & Bekelman, T. A. (2024). Socioeconomic status and obesity. *Journal of the Endocrine Society, 8*(11), bvae176. https://doi.org/10.1210/jendso/bvae176

Centers for Disease Control and Prevention. (2022, July 15). *Consequences of obesity.* National Center for Chronic Disease Prevention and Health Promotion, Division of Nutrition, Physical Activity, and Obesity. https://www.cdc.gov/obesity/basics/consequences.html

Centers for Disease Control and Prevention. (2024 April 2). *Childhood obesity facts.* U.S. Department of Health and Human Services. https://www.cdc.gov/obesity/childhood-obesity-facts/childhood-obesity-facts.html

Centers for Disease Control and Prevention. (2024, May 14). *Adult obesity facts.* U.S. Department of Health and Human Services https://www.cdc.gov/obesity/adult-obesity-facts/

Centers for Disease Control and Prevention. (2024, September 12). *Adult obesity prevalence maps* [Web page]. U.S. Department of Health and Human Services. Retrieved from https://www.cdc.gov/obesity/data-and-statistics/adult-obesity-prevalence-maps.html

Centers for Disease Control and Prevention. (2025, January 30). *Nutrition, physical activity, and obesity — Behavioral Risk Factor Surveillance System* [Dataset]. Data.gov. Retrieved September 16, 2025, from https://catalog.data.gov/dataset/nutrition-physical-activity-and-obesity-behavioral-risk-factor-surveillance-system

Centers for Disease Control and Prevention. (April 21, 2025). *Obesity among children and adolescents aged 2–19 years, by selected characteristics: United States* [Dataset]. Data.gov. Retrieved from

https://data.cdc.gov/National-Center-for-Health-Statistics/Obesity-among-children-and-adolescents-aged-2-19-y/9gay-j69q

Capoccia, D., Milani, I., Colangeli, L., Parrotta, M. E., Leonetti, F., & Guglielmi, V. (2025). Social, cultural and ethnic determinants of obesity: From pathogenesis to treatment. Nutrition, Metabolism & Cardiovascular Diseases

Cohen, A. K., Rai, M., Rehkopf, D. H., & Abrams, B. (2013). Educational attainment and obesity: A systematic review. *Obesity Reviews, 14*(12), 989–1005. https://doi.org/10.1111/obr.12062

Economic Research Service, U.S. Department of Agriculture. (2025, January 5). *Food Access Research Atlas – Download the data* [Dataset]. Retrieved from https://www.ers.usda.gov/data-products/food-access-research-atlas/download-the-data

Estimation of Obesity Levels Based on Eating Habits and Physical Condition [Dataset]. (2019). *UCI Machine Learning Repository.* https://doi.org/10.24432/C5H31Z

GBD 2021 US Obesity Forecasting Collaborators. (2024). National-level and state-level prevalence of overweight and obesity among children, adolescents, and adults in the USA, 1990–2021 and forecasts up to 2050. *The Lancet, 404*(10469), 2278–2298. https://doi.org/10.1016/S0140-6736(24)01548-4

Geomontes. (n.d.). *Obesity, poverty, and income in U.S. (2019–2023)* [Dataset]. Kaggle. Retrieved from https://www.kaggle.com/datasets/geomontes/obesity-poverty-and-income-in-u-s-20192023

GlobalData Healthcare (2023, February 17). Racial and socioeconomic disparities increase childhood obesity in the US. GlobalData Healthcare. https://www.clinicaltrialsarena.com/analyst-comment/disparities-childhood-obesity/?cf-view

Kelly, C. N., Da Rosa, P., Remmele, J., Tilenbaeva, N., Arnold, R., Oberhoffer-Fritz, R., Fonseca, H., Buoncristiano, M., Wickramasinghe, K., & Williams, J. (2025). Inequalities in childhood overweight

and obesity: A call to strengthen upstream policy measures. *Public Health in Practice, 10,* 100637. https://doi.org/10.1016/j.puhip.2025.100637

Montes, G. (2023). Obesity, poverty, and income in U.S. (2019–2023) [Dataset]. Kaggle. https://www.kaggle.com/datasets/geomontes/obesity-poverty-and-income-in-u-s-20192023 kaggle.com+1

National Center for Health Statistics. (n.d.). *Obesity among children and adolescents aged 2–19 years, by selected characteristics: United States*. Centers for Disease Control and Prevention. https://data.cdc.gov/d/9gay-j69q

Our World in Data: Share of children and adolescents who are overweight or obese." (2025). Adapted from World Health Organization data. Retrieved October 5, 2025, from https://ourworldindata.org/grapher/child-adolescent-obesity

Polyzou, E. A., & Polyzos, S. A. (2024). Outdoor environment and obesity: A review of current evidence. *Metabolism Open, 24,* 100331. https://doi.org/10.1016/j.metop.2024.100331

Randolph, J., & Stephens, J. (2022, January 10). *Social determinants of health and obesity: Implications for clinical research and practice.* Preventive Cardiovascular Nurses Association (PCNA). https://pcna.net/news/social-determinants-of-health-and-obesity/

Sanyaolu, A., Okorie, C., Qi, X., Locke, J., & Rehman, S. (2019). Childhood and adolescent obesity in the United States: A public health concern. *Global Pediatric Health, 6*, 2333794X19891305. https://doi.org/10.1177/2333794X19891305

Stohl, E. (2023, August). Child obesity in the United States. Ballard Brief. https://ballardbrief.byu.edu/issue-briefs/childhood-obesity-in-the-united-states

Tcrammond. (n.d.). *Food access and food deserts* [Dataset]. Kaggle. Retrieved from https://www.kaggle.com/datasets/tcrammond/food-access-and-food-deserts

UoM190346A. (n.d.). *Sleep Health and Lifestyle Dataset* [Dataset]. Kaggle. Retrieved from https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset

U.S. Centers for Disease Control and Prevention. (n.d.). *National obesity prevalence by state* [Dataset]. Data.gov. Retrieved from https://catalog.data.gov/dataset/national-obesity-by-state-d765a

Zhang, Z., Xu, Z., Zhou, L., Yang, J., Wang, J., & Yang, Q. (2024). *Unsupervised anomaly detection for multi-stream time series based on large language models*. **Frontiers in Big Data, 7**, 1469981. https://doi.org/10.3389/fdata.2024.1469981