### Trends in Obesity Rates Among High School Students in the U.S. from 2011 to 2021: Associations with Soda Consumption and Physical Activity Across Racial and Ethnic Groups

**Klarissa Siebert**  
Department of Public Health, The University of Texas at El Paso  
PUBH 5357: Health Data Science - Storytelling with Data  
**Dr. Amy Wagler**  
May 16, 2025

**Research Question**

How have obesity rates among high school students in the U.S. changed from 2011 to 2021, and how are these trends associated with changes in soda consumption and physical activity prevalence across racial and ethnic groups?

**Brief Summary of Results**

Obesity prevalence increased significantly from 2011 to 2021 among U.S. high school students. Additionally, race/ethnicity was significantly associated with obesity, particularly for Hawaiian/Pacific Islander students, who had the highest rates. However, soda consumption and physical activity were not found to be statistically significant predictors of obesity. The findings from this research suggest that the increase in obesity appears to be influenced more by time and racial disparities rather than behavioral factors alone.

# Introduction

Obesity among children and adolescents in the United States continues to be a public health issue. National estimates show that nearly 20% of youth aged 2 to 19 years were classified as obese in 2015 to 2016, with even higher rates among the adolescent group (Sanyaolu et al., 2019). Obesity is important to address in children because it increases the risk of chronic diseases such as type 2 diabetes, cardiovascular disease, and certain cancers later in life (Sanyaolu et al., 2019). Disparities in obesity rates are also prevalent across racial and ethnic groups, with non-Hispanic Black and Hispanic adolescents showing significantly higher prevalence than their White counterparts (Singh et al., 2008). In addition to racial disparities, physical activity has been associated with higher obesity prevalence among youth, with sedentary children experiencing significantly greater odds of obesity (Sing et al., 2008).

Youth obesity rates have continued to rise over the past two decades. Wang et al. (2020) found a steady increase in obesity and severe obesity among adolescents from 1999 to 2016 and projected that nearly half of U.S. adolescents may be overweight or obese by 2030. They attribute this trend to a combination of poor diet, inadequate physical activity, and socioeconomic inequalities. Research from outside the U.S. supports this idea as well. Rizwan et al. (2011), who studied adolescents in Pakistan, found strong links between sugar-sweetened beverage consumption, insufficient school physical activity, and increased obesity. Their findings support the idea that school and environmental factors influence youth health behaviors and outcomes.

Given the public health importance of decreasing adolescent obesity and the known behavioral influences on weight, this study investigates obesity trends among U.S. high school students using nationally representative surveillance data. Specifically, I examine how obesity prevalence changed from 2011 to 2021 and whether these trends are associated with soda consumption and physical activity levels across racial and ethnic groups. This analysis aims to understand whether modifiable health behaviors explain disparities in youth obesity or whether demographic and structural factors play a stronger role.

This project is intended for public health professionals, school health policy advocates, and community health researchers interested in adolescent obesity prevention. The findings may also be useful to school administrators and public health departments working to reduce behavioral and racial disparities in youth health outcomes using population-level data.

# Methodology

To answer my research question, I used data from the Youth Risk Behavior Surveillance System (YRBSS), which I retrieved from the website Data.Gov in CSV format (2020). I cleaned the dataset to contain only variables that were related to obesity prevalence, daily soda consumption, physical activity levels, racial/ethnic identity, and year of data collection. The dataset was also cleaned to only contain entries that had been previously stratified by race/ethnic category to prevent missing data. The data was then organized by year and racial/ethnic group to assess trends over time and between populations. Data visualizations were created to explore variable distributions and temporal trends (see Appendix). Also, a heat map was created and used to assess multicollinearity among the variables (see Appendix).

## *Regression Models and Assumptions*

A simple linear regression was run to examine the relationship between year and obesity prevalence. A multiple linear regression model was then used to explore how soda consumption, physical activity, race/ethnicity, and year were associated with obesity prevalence. I assessed multicollinearity in the multiple regression model using Generalized Variance Inflation Factors (GVIFs), which were all below 3.0, indicating no major concerns. The highest GVIF was 2.87 for soda consumption, followed by 2.03 for physical activity, 2.07 for year, and 1.23 for race/ethnicity. Residuals from both models were visualized using histograms and quantile-quantile plots and were approximately normally distributed, meaning that both models met the assumptions for linear regressions.

# Process and Tools

The entire analysis was conducted in Google Colab, a cloud-based coding platform. I used Gemini, the AI assistant, within Google Colab to help generate R code and create visualizations. The raw dataset was first cleaned in Microsoft Excel and then uploaded to Google Colab for analysis. Once the analysis was complete, all scripts, visualizations, and outputs were saved to a GitHub repository for future use.

# Statistical Approach

Descriptive statistics were calculated for the key predictors, including obesity prevalence, soda consumption, and physical activity levels (see Table 1). Mean values of health indicators by race and year were also calculated (see Appendix). Also, a correlation matrix was created to explore any relationships between the variables.

**Table 1.** **Descriptive Statistics for U.S. High School Students (n=175)**

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistic** | **Obesity Prevalence (%)** | **Soda Consumption (%)** | **Physical Activity (%)** |
| **Min** | 5.50 | 4.60 | 15.30 |
| **Q1** | 12.18 | 15.15 | 21.45 |
| **Median** | 15.65 | 19.70 | 26.15 |
| **Mean** | 15.51 | 19.76 | 25.68 |
| **Q3** | 18.27 | 23.90 | 28.23 |
| **Max** | 29.40 | 35.80 | 40.00 |

## *Modeling Strategy*

A simple linear regression model was used to evaluate trends in obesity over time. Then, a multiple linear regression model was run to examine how soda consumption, physical activity, race/ethnicity, and year were associated with obesity prevalence. Visualizations, including scatter plots, histograms, and line graphs, were used to illustrate variable trends over time and by racial/ethnic group (see Appendix).

# Analysis and Results

The simple linear regression showed that obesity rates significantly increased over time (see Table 2). This model confirmed this trend by using year as a predictor (Estimate = 0.5635, p = 0.0314, 95% CI [0.053, 1.074]) and also explained 9.3% of the variance in obesity (Adjusted R² = 0.093). Residuals from this model were approximately normally distributed, which verifies the validity of the results.

**Table 2.** **Simple Linear Regression Results Predicting Obesity Over Time**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor** | **Estimate** | **Std. Error** | **p-value** | **95% CI** |
| Intercept | -1120.43 | 508.49 | 0.0337\* | [-2149.81, -91.04] |
| Year | 0.5635 | 0.2522 | 0.0314\* | [0.0529, 1.0741] |

**Note:** Adjusted R² = 0.12. \* denotes a significant value of < 0.05.

The multiple linear regression model revealed several key findings (see Table 3). Soda consumption and physical activity were not significantly associated with obesity. However, the students who identified as Hawaiian/Pacific Islander showed a significant positive association with obesity prevalence. The year of data collection was also a significant predictor, suggesting that obesity has increased over time regardless of behavior changes. The full model explained about 57% of the variance in obesity (Adjusted R² = 0.571). A scatterplot with a regression line and a forest plot were created to visualize the main model (see Appendix).

**Table 3.** **Multiple Linear Regression Results Predicting Obesity Prevalence**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Predictor** | **Estimate** | **Std. Error** | **p-value** | **95% CI** |
| Intercept | 1695.65 | 738.77 | 0.029\* | [-3206.60, -184.69] |
| Soda Consumption | 0.101 | 0.232 | 0.666 | [-0.374, 0.576] |
| Physical Activity | 0.380 | 0.231 | 0.111 | [-0.093, 0.854] |
| Hawaiian/Pacific Islander | 5.363 | 2.496 | 0.040\* | [0.259, 10.467] |
| Non-Hispanic White | -4.274 | 2.269 | 0.070 | [-8.914, 0.367] |
| Year | 0.843 | 0.363 | 0.028\* | [0.100, 1.587] |

**Note:** Only selected predictors shown. Full model F(9, 29) = 6.618, p < 0.001.

Adjusted R² = 0.571. \* denotes a significant value of < 0.05.

# Discussion

This study found that obesity rates among high school students in the U.S. have significantly increased from 2011 to 2021. This finding is supported by both the simple regression model and the upward trends seen in the visualizations. The Hawaiian/Pacific Islander group had the highest predicted obesity rates, a finding that was statistically significant even after adjusting for other soda consumption, physical activity, and year. Hispanic and Black youth also had consistently elevated average obesity rates, though not statistically significant in the main regression model.

Although soda consumption and physical activity are commonly known in the literature as contributors to obesity, they were not significant predictors in this dataset. This could be due to self-reported data bias, measurement error, or the outside influence of other factors not measured in the statistical models. However, it is important to mention that soda consumption declined over time across all groups, while obesity continued to rise. Physical activity rates also declined, except among American Indian/Alaska Native students. These trends suggest that broader environmental, structural, and demographic influences, such as poverty, access to healthy food, school funding, racial discrimination, and community resources, may play a stronger role than individual behaviors.

***Limitations***

One limitation of this project is the reliance on self-reported survey data, which can introduce bias. Since the high school students completed the survey themselves, that may have posed issues such as understanding the questions, recalling their behaviors correctly, or putting the answers that they believe are most desired even if they are untrue or inaccurate. Additionally, the YRBSS dataset may not fully represent all students, especially in states or districts with lower participation rates. These gaps can hide the experiences of underrepresented groups and limit the ability to draw conclusions that are equitable. From a data equity perspective, the lack of disaggregated data within certain racial or ethnic subgroups may also mask important differences, reinforcing the need for more inclusive data collection methods in student surveys.

# Data Equity

As previously mentioned, when analyzing trends by race and ethnicity, it is important to consider data equity principles. As outlined by the Public Health Institute at Denver Health (2024) Data Equity Principles Guide, race should be considered a social construct, not a biological one. This perspective helps prevent generalizations about certain individuals or communities and avoids placing any blame on them. Also, disaggregating data by race and ethnicity provides a clearer understanding of health disparities. It is essential to acknowledge the limitations in data collection, data analysis, and methodology to maintain transparency and trust.

# FAIR Principles

The YRBSS dataset follows many of the FAIR data principles (GO FAIR, 2016). It is Findable through Data.Gov and Accessible in multiple formats compatible with several statistical tools. The dataset is Interoperable with programs like SPSS, SAS, and R. However, its Reusability could be improved with the inclusion of a codebook for users to better understand the dataset. Additionally, variations in state participation may impact the generalizability of the findings.

# Conclusion

This project contributes to our understanding of youth obesity by showing that rates have continued to increase over the past decade and that disparities persist across racial and ethnic groups. The analysis found that while soda consumption and physical activity levels are important public health indicators, they were not significant predictors of obesity in this dataset. Instead, the year of data collection and racial identity, particularly for Hawaiian/Pacific Islander youth, were stronger predictors.

These findings emphasize the need for culturally relevant interventions that address not only student behaviors but also the broader environmental factors affecting adolescent health. Examples if interventions include school nutrition and physical activity programs tailored to specific communities, improved access to safe spaces for physical activity, and advocating for equitable school funding and food. Future research should explore how structural factors may contribute to obesity disparities.

### References

GO FAIR. (2016). FAIR Principles Overview. Retrieved from GO FAIR.

Public Health Institute at Denver Health. (2024). PHIDH Data Equity Principles Guide. Denver, CO: Public Health Institute at Denver Health.

Rizwan, F., Qadir, M. A., Hussain, H., & Hameed, A. (2011). Prevalence of obesity and its relation to lifestyle behaviors in school children of Lahore. *Annals of King Edward Medical University*, *17*(2), 129–133. <https://doi.org/10.21649/akemu.v17i2.109>

Sanyaolu, A., Okorie, C., Qi, X., Locke, J., Rehman, S., Patidar, R., & Prakash, S. (2019). Childhood and adolescent obesity in the United States: A public health concern. *Global Pediatric Health*, *6*, 1–11. <https://doi.org/10.1177/2333794X19891305>

**Singh, G. K., Kogan, M. D., & Van Dyck, P. C. (2008).** A multilevel analysis of state and regional disparities in childhood and adolescent obesity in the United States. Journal of Community Health, 33(2), 90–102. <https://doi.org/10.1007/s10900-007-9071-7>

U.S. Department of Health & Human Services. (2020, November 10). *Nutrition, Physical Activity, and Obesity - Youth Risk Behavior Surveillance System*. data.gov. <https://catalog.data.gov/dataset/nutrition-physical-activity-and-obesity-youth-risk-behavior-surveillance-system>

Wang, Y. C., Beydoun, M. A., Liang, L., Caballero, B., & Kumanyika, S. K. (2020). Will all Americans become overweight or obese? Estimating the progression and cost of the US obesity epidemic. *Obesity*, *18*(10), 1904–1912. <https://doi.org/10.1038/oby.2010.100>

**Appendix**

**A. Mean Values of Health Indicators by Race**

|  |  |  |  |
| --- | --- | --- | --- |
| **Race/Ethnicity** | **Obesity (%)** | **Soda Consumption (%)** | **Physical Activity (%)** |
| **2 or more Races** | 16.35 | 19.28 | 26.87 |
| **American Indian/Alaska Native** | 17.37 | 29.26 | 32.93 |
| **Asian** | 7.17 | 9.23 | 18.43 |
| **Hawaiian/Pacific Islander** | 21.03 | 18.13 | 26.48 |
| **Hispanic** | 17.22 | 19.73 | 23.70 |
| **Non-Hispanic Black** | 18.53 | 22.40 | 23.63 |
| **Non-Hispanic White** | 12.72 | 21.35 | 28.00 |

**B. Mean Values of Health Indicators by Year**

|  |  |  |  |
| --- | --- | --- | --- |
| **Year** | **Obesity (%)** | **Soda Consumption (%)** | **Physical Activity (%)** |
| **2011** | 15.16 | 26.31 | 27.33 |
| **2013** | 11.63 | 24.64 | 26.73 |
| **2015** | 14.08 | 19.80 | 27.32 |
| **2017** | 16.03 | 17.90 | 26.27 |
| **2019** | 16.13 | 13.24 | 21.82 |
| **2021** | 19.89 | 14.83 | 24.31 |

**C. Histogram Distributions of Key Variables**

**A graph of a number of blue bars

AI-generated content may be incorrect.**

**A graph of a physical activity

AI-generated content may be incorrect.**

**A graph of a distribution of soda consumption

AI-generated content may be incorrect.**

**D. Health Indicators by Race/Ethnicity – Line Graphs**

**A graph with colored lines and numbers

AI-generated content may be incorrect.**

**A graph of different colored lines

AI-generated content may be incorrect.**

**A graph of the number of soda consumption trends

AI-generated content may be incorrect.**

**E. Correlation Matrix – Heat Map of Predictors**

**A graph with different colored squares

AI-generated content may be incorrect.**

**F. Scatterplot with Regression Line – Simple Linear Regression: Obesity Over Time**

**A graph with a line and dots

AI-generated content may be incorrect.**

**G. Forest Plot of Multiple Linear Regression Results**

**A graph of obesity

AI-generated content may be incorrect.**

**A blue graph with black background

AI-generated content may be incorrect.H. Normality Test for Simple Linear Regression Model**

**A red line in a black background

AI-generated content may be incorrect.**

**A blue rectangular object with black background

AI-generated content may be incorrect. I. Normality Test for Multiple Linear Regression Model**

**A red line in a black background

AI-generated content may be incorrect.**