Analysis of the correlation between diabetes rates and access to healthy food

Ngoc Bao Anh Le (ngocbaoanh.le@concordiahanoi.org)
Big Data 1 2025-2026 (Ms. Thu Hien Le)

Introduction

Public concern regarding the health of younger generations in the United States has intensified in recent years. In particular, the prevalence of diabetes has risen at an alarming rate, especially among younger age groups, posing a significant public health challenge. The United States, despite being one of the world's leading nations in agriculture and food production, has exhibited rising diabetes rates in recent years. This has drawn the public's attention to concerns about the rise of health problems in the future. This study aims to investigate whether insufficient access to healthy food correlates with a higher rate of diabetes across the U.S. More specifically, this study explores the relationship between the population with insufficient access to healthy food options in the U.S. and the percentage of the population in different states with diabetes using data from the USDA Food Environment Atlas, the National Center for Health Statistics, and Kaggle.

Data collection

There are three data sets used in this study. The first data set is collected from the U.S. Department of Agriculture Food Environment Atlas as an Excel file. The two sections primarily used in this data set are Health and Access. The second data set is collected from the National Center for Health Statistics as a csv file. The data primarily used in this data set is the numerical characterization based off the degree of urbanization of areas in the U.S. ranging from 1, large central metro, to 6, noncore. The third data set is collected secondarily from Kaggle about U.S. county-level demographic, economic, and health information. The data primarily used in this data set is the median household income.

Data Analysis

A. The correlation between diabetes rates and access to healthy food

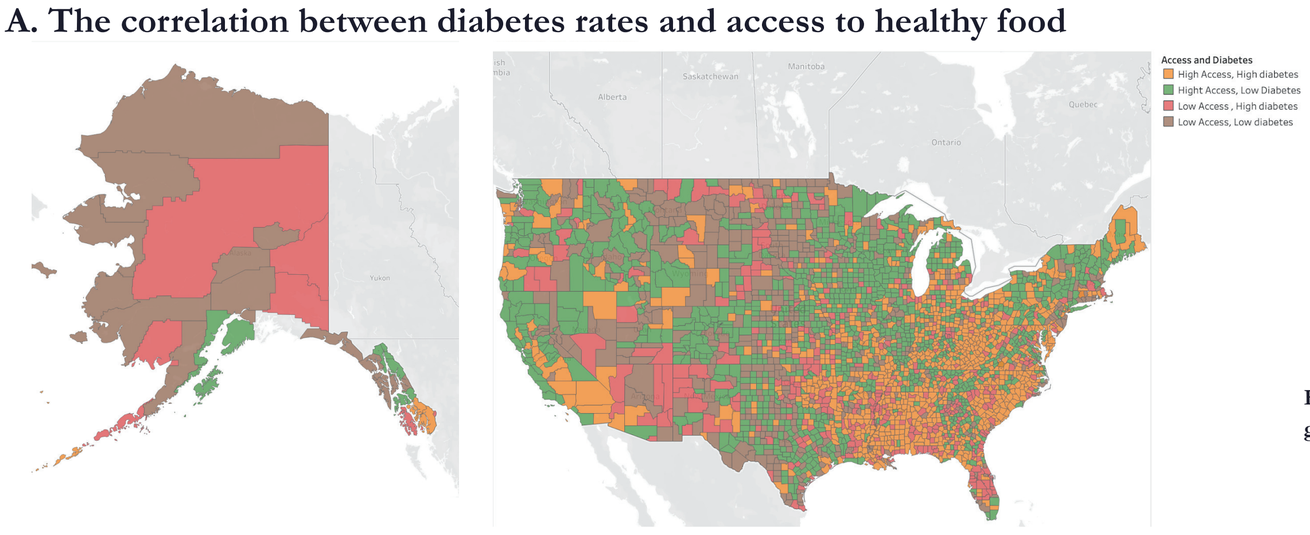

Figure 1. Geographic Distribution of Food Access and Diabetes Rates in the United States

Based on the geographical trends depicted in Figure 1, two areas are selected to compare the percentages and number of grocery stores per 1000 population. The two areas chosen are Appalachian and Southern and Western and Northeastern. The analysis shows that the largest proportion of group 1 is "high access, high diabetes" and the largest proprotion of group 2 is "high access, low diabetes". Notably, the comparison the average number of grocery stores per 1000 population of the two areas shows that group 2 have significantly higher number of average grocery stores than group 1(Figure 2). Synthesizing the results of the two comparisons, group 2 with a higher number of average grocery stores exhibit lower diabetes rates than group 1. Additionally, the graph exhibits the relationship between the number of grocery stores and diabetes rates, which showed a weak to moderate correlation with an r^2 of 0.32102(Figure 3).

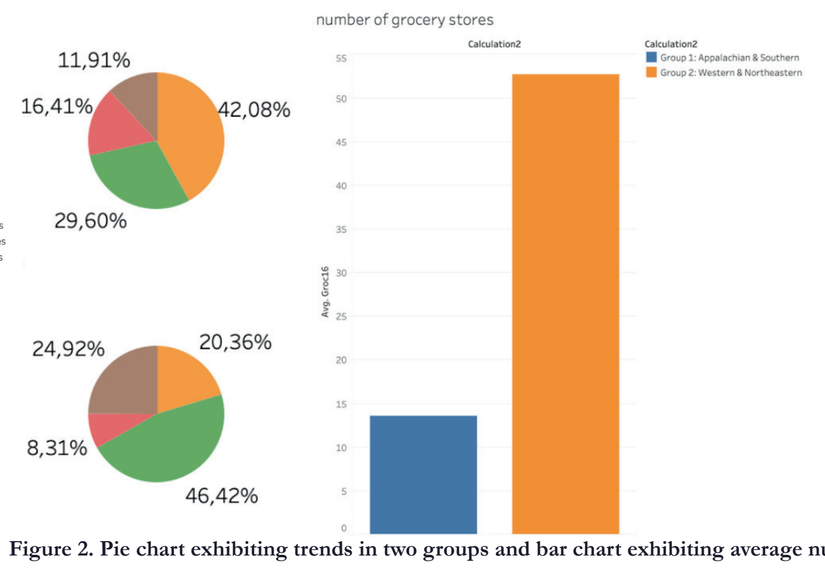

Figure 2. Pie chart exhibiting trends in two groups and bar chart exhibiting average number of grocery stores in 2 groups.

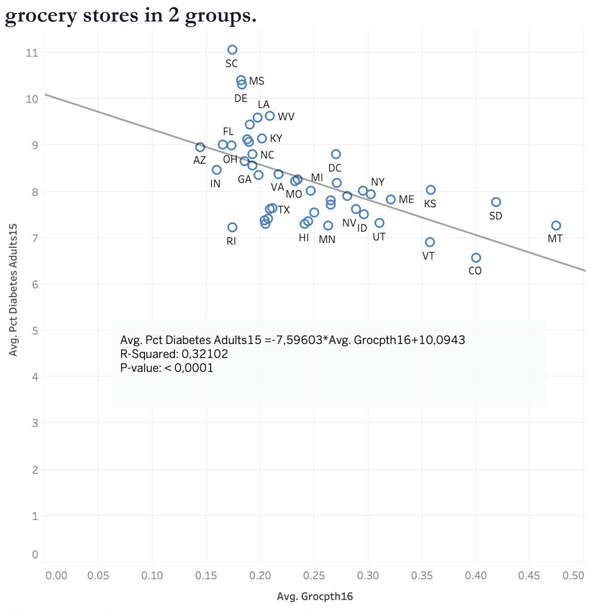

Figure 3. Correlation between diabetes rates and the number of grocery stores per 1000 population

B. The correlation between diabetes rates and other confounding variables

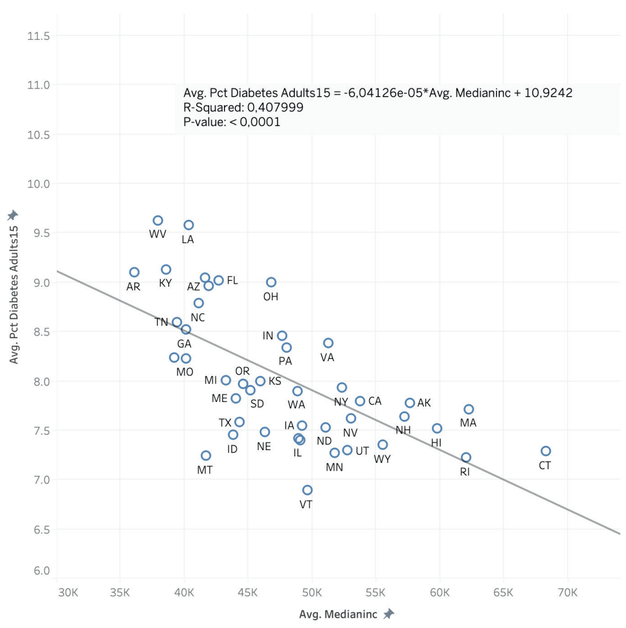

Figure 4. Correlation between diabetes rates and median household income

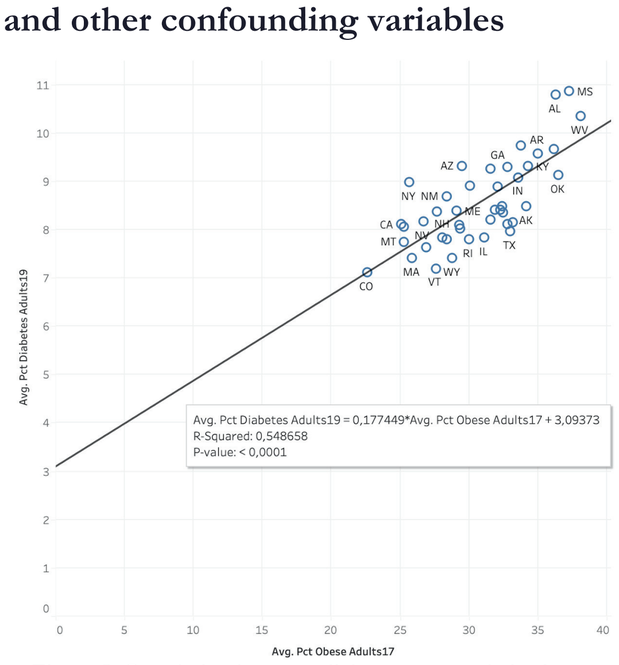

Figure 5. Correlation between diabetes rates and obesity rates

Analysis of the correlation between diabetes rates and median household income showed an r2 value of 0.407999, which hints at a moderate relationship between diabetes rates and median household income (Figure 4). Median household income is chosen as a confounding variable because income plays a decisive role in financial decisions of households. Therefore, median household income is used to analyze its relationship with diabetes rates. Additionally, analysis of the correlation between diabetes rates and obesity rates shows an r2 value of 0.548658, which shows a moderate relationship between obesity and diabetes (Figure 5). These two correlations of diabetes rates and two confounding variables show that the relationship between diabetes rates and other confounding variables is significant. Notably, the relationship between diabetes rates and base health factor shows the largest r2 value across all analyses.

Discussions and Limitations

In summary, all analyses show that the relationship between diabetes rates and access is significant but non-exclusive; rather, diabetes rates are more correlated with other confounding variables, median household income and base health issue. Firstly, the comparison of the two groups shows that the group with a larger number of grocery stores exhibit lower diabetes rates, but the correlation between diabetes rates and number of grocery stores shows a moderate to weak relationship with an r^2 value of 0.32102. In contrast, both analyses of the correlation between diabetes rates with median household income and obesity rates show r^2 values of 0.407999 and 0.548658, respectively. This shows that both the correlations between diabetes with the two confounding variables have a stronger correlation with diabetes rates than access to healthy food does. Overall, the correlation between diabetes rates and access to healthy food is significant but non-exclusive.

The study relies primarily on data from 2013–2019 (notably 2015), limiting its contemporary applicability. Additionally, the datasets are temporally misaligned—for example, obesity data from 2017 were paired with diabetes data from 2019—which may reduce analytical accuracy. Finally, because the study depends on secondary data from the USDA and NCHS that include substantial self-reported measures, the results may be affected by reporting bias. Addressing these issues would strengthen future research.

Future Work

Furthermore, this research aspires to connect the results of the analysis to the broader global context of health inequality. By examining patterns across states, this study hopes to provide insights into how insufficient healthy food access may contribute to the worldwide rise in chronic health problems.

Conclusion

The study demonstrated the relationship between diabetes rates with access to healthy food and the relationship between diabetes rates with confounding variables focusing from 2013-2019. Analysis shows that the relationship between diabetes rates and access to healthy food is significant but non-exclusive; rather, diabetes rates are more strongly correlated with socioeconomic factors such as income and base health issues. For future work, the correlation between access to healthy food and diabetes rates can be more accurately measured by precisely controlling confounding variables.

Full paper