Identifying the Relationship Between Food Access & Diabetes Prevalence by County in the United States

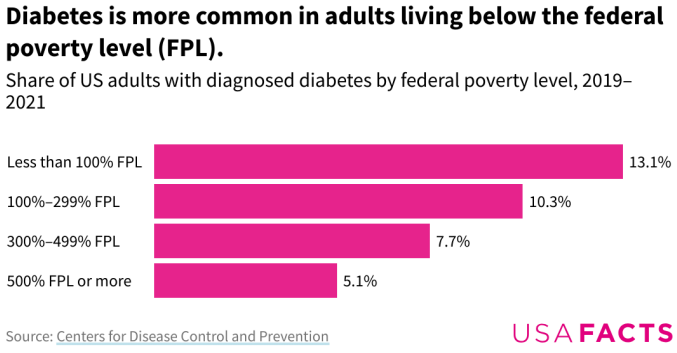
Ryan Manthy
Illinois Institute of Technology
rmanthy@hawk.iit.edu

Abstract and Problem Statement

In the United States, 38.4 million Americans are diagnosed with Diabetes. This number has grown in the last 20 years by more than 26%. Additionally, an estimated 97.6 million American adult have pre-diabetes, a condition that is an early indicator that they may develop diabetes in the near future.

Type II Diabetes – the condition growing at the greatest rate – is a lifestyle disease, meaning that the dietary and exercise choices made by individuals influences their likelihood in developing the condition. While lifestyle choices may be voluntary, the conditions around access to healthy food often may limit peoples' abilities to make choices in the best interests of their health.

Individuals living below the poverty level are far more likely to be diagnosed with Diabetes than their counterparts living above the poverty level. In particular, there is a strong negative correlation between income and diabetes likelihood. (Source: USA Facts)



Working together with lower income levels, food insecurity can also make it difficult for people to have diets that avoid the processed and saturated foods that increase the likelihood of diabetes in adults.

This research study seeks to better understand the relationship between food access and Type II diabetes by evaluating the trends of both at a county-level and determining if there is a statistically significant relationship between the two.

Data Selection

In identifying appropriate datasets for this analysis, the research focused on primary sources, including recent available government data on the subjects. Food insecurity is a difficult trend to measure at a county-level, so the analysis must focus on quantitative metrics that can be used to confirm or fail to confirm a statistically significant relationship between food access and diabetes prevalence.

The study, thus, focuses on two datasets: the USDA's Food Access Research Atlas and the CDC's Diabetes atlas for evaluation. Each aggregates data points and indicators that can be used in the analysis of their respective domains.

Food Access Research Atlas

The Food Access Research Atlas specifically emphasizes data points that indicate food access, including SNAP/EBT participation, supermarket accessibility rates, and household income. These indicators are brought together to develop **Low-income and Low-access (LILA)** numbers for areas of greatest identified need. LILA specifically has counts for both rural and urban areas.

These data points are accessible at the census tract level – a level of greater granularity than the CDC's Diabetes Atlas. Data in the latest report is from 2019, so this will be used in our analysis. It is available in a CSV format and is machine readable.

There are 72,532 Census Blocks in the United States each representing a relatively small unit of population.

CDC Diabetes Atlas

The CDC developed a Diabetes Surveillance System that is accessible at the county level and displays data on the rate of *diagnosed* diabetes and obesity as a percentage of the population. The most recent available Diabetes data is from 2021.

Diabetes data isn't in itself useful, so it is helpful to have additional data to build a frame of reference compared to this provided CDC information. The CDC also provides Social Determinants of Health (SDOH) data that can provide important context into the ways racial, gender, and ethnic-based healthcare access gaps impact diabetes prevalence.

There are 3,078 recorded counties in the CDC's dataset.

Data Integration & Pre-Processing

Having data in different structures between the Diabetes Atlas and the Food Access Research Atlas makes analysis difficult without pre-processing. To integrate the data appropriately, it is necessary to organize the Census Blocks by county and validate that both datasets are of equal granularity and organization.

This data integration can be done using Python and will provide us with a robust information repository that can be used to generate statistically significant predictions about the data.

Completing the data integration, we find that there are 3,143 unique counties in the Food Research Atlas compared to 3,078 counties in the Diabetes Atlas. For the sake of this exercise, if there wasn't a present match, the county was thrown out.

In the pre-processing, the delineation made was if greater than 50% of Census Blocks in a county were a designated LILA track, the county was given a score of 1. Otherwise, the score was 0. This does present a binary view of food insecurity, but it should be sufficient for determining if there is a potential correlation.

Food Research Atlas	Diabetes Atlas
3,143 Counties	3,078 Counties

Hypothesis Generation

Based on the preliminary research conducted around the known relationship between poverty and diabetes, the hypotheses generated will assume that food insecurity does correlate with diabetes rates at a county-level.

These county-level predictions may also be used to inform predictions at a Census Block level. However, we lack the data to conduct a study on Census Block data.

Key Variables

Independent Variables

LILATracts_1AND10 – These are Census Block tracts where the median income is considered low income and a significant percentage of the population is more than 1 mile (urban) or 10 miles (rural) from a supermarket.

Dependent Variables

Diabetes Prevalence – This data is collected at the county level and highlights the percentage of the population diagnosed with diabetes.

Covariates

Obesity Rate – This data is collected at the county level and highlights the percentage of the population diagnosed with obesity.

Poverty Rate – This data is collected at the Census Block level and highlights the percentage of the population at or below the poverty threshold.

Hypothesis

The identified variables will be used to generate a primary and secondary hypothesis about the relationship between diabetes and food access.

Primary Hypothesis

Our primary hypothesis is that counties identified as food insecure (formerly known as food deserts) will have a higher diabetes rate than food secure counties.

$$H_0: \mu_{diabetes,\, food\ insecure} \leq \mu_{diabetes,\, food\ secure}$$
$$H_A: \mu_{diabetes,\, food\ insecure} > \mu_{diabetes,\, food\ secure}$$

This hypothetical relationship is represented above as a null hypothesis that the average diabetes rate for food insecure counties is less than or equal than a food secure county and vice-versa for the alternative hypothesis.

Secondary Hypothesis
Our secondary hypothesis is that poverty rates at a county-level correlate with the diabetes prevalence at a county-level.

$H_0: \rho \leq 0$
$H_A: \rho > 0$

ρ is the correlation coefficient between PovertyRate and Diagnosed Diabetes (as a percentage).

Study Construction

Primary Hypothesis Study

To test the hypothesis we will need the binary measure of food insecurity or security and diabetes prevalence at a county-level. This data is attached to CountyID in the processed data. Because the data is generated from census data, the sample and target populations are effectively the same and sampling bias is not a strong consideration. The only important factor to consider is that the data is now between 3 and 5 years old, so its dated nature may make it distinct from the target population.

The required pre-processing was already completed at the pre-processing stage. All 3,078 available counties will be used in the study to generate the most complete data. This size of data far exceeds any minimum sample size for a study like this and will provide statistically significant findings at α=0.05.

To control for extraneous effects, PovertyRate and Obesity will be provided as control variables. However, it is worth noting that these variables are likely covariates to both diabetes and food insecurity.

Secondary Hypothesis Study

The secondary hypothesis will be tested using PovertyRate and Diagnosed Diabetes as a percentage of the population at a county level. As was the case in the primary hypothesis, the secondary hypothesis is based on Census data and so there is little need to account for sampling bias or differences between target and sample populations.

The required pre-processing was also completed when the data was cleaned from the target data sources.

To account for potential covariates in the secondary hypothesis study, multiple linear regression can be used to identify how PovertyRate interacts with obesity and food insecurity data in the findings.

Conducting Study & Analysis

Primary Hypothesis Study

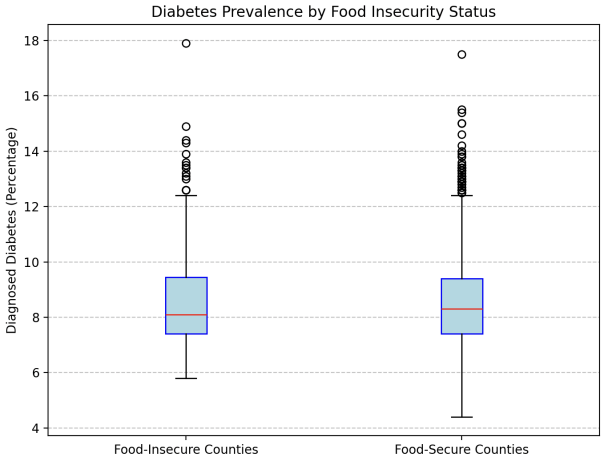
To generate a study and analysis for the Primary Hypothesis, two statistical tests will be employed. The results from these tests will be compared at the end to verify accuracy and consistency.

Independent Samples t-Test (one-tailed)
The first test for food insecurity and diabetes is a t-Test. Because the samples are independent and this test is fairly straightforward to use in beginning the investigation process. We will break the data into two groups: LILATracts = 1 and LILATracts = 0.

Running the study across all 3,078 counties, the results ***failed to reject the null hypothesis***. This suggests that there is no significant difference in diabetes prevalence between food secure and food insecure counties.

t-Statistic: 0.9533
p-Value: 0.1703

These results prompt further investigation, which by using a box and whiskers plot reveal that while there are more counties with exceptionally low diabetes rates in the food secure category, the middle range and averages are effectively the same.



Pearson Correlation Coefficient
To validate the information from the plots and t-Test, it is helpful to use another statistical measure. To verify this, an investigation of various similarity metrics led to the Pearson Correlation Coefficient being a suitable measure for diabetes prevalence similarity.

The large number of counties (> 3,000) means that we can assume the values are normally distributed.

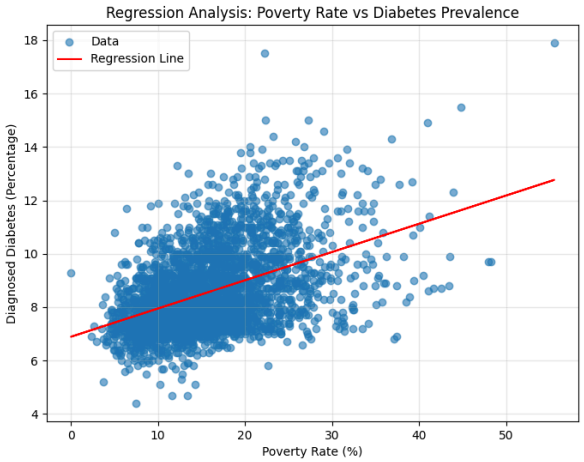
Running the Pearson Correlation Coefficient test reveals the same findings as the t-Test: we ***fail to reject the null hypothesis***.

Pearson Correlation (r) = 0.0175
p-Value = 0.3405

Secondary Hypothesis Study

The secondary hypothesis generated for this study suggests a relationship between the poverty rate and rate of diagnosed diabetes by county. This initial hypothesis was generated based on the known relationship between poverty and diabetes prevalence. Studying the relationship at a county level seeks to reject or fail to reject that a higher poverty rate leads to higher diabetes rates.

Simple Linear Regression
For the secondary hypothesis, using simple linear regression will help to quantify – if a relationship does exist – what the rate of change is in comparison between the two data points. This makes sense for something like comparing the county obesity levels to the diabetes prevalence.



Conducting a simple linear regression between poverty rate and diabetes prevalence reveals a clear correlation at the county level.

Variable	Coefficient	Std. Error	t-value
Diabetes	6.893	0.0665	103.6167
PovertyRate	0.1057	0.0039	27.19419

Analyzing these regression results reveals that there exists a relationship between poverty and diabetes prevalence at the county level. The regression model states that for every 1% increase in the poverty rate, there is a 0.1057% increase in diabetes.

Additionally, the remaining elements of the data point to this being a statistically significant claim. The Standard Error of 0.0039 suggests a precise estimate, a high t-Value of 27.194 indicates significant statistical difference from 0, and the P-Value is very small which suggests ***rejecting the null hypothesis.***

Discussion

The results of this study varied from the expected result in that the **primary hypothesis failed to be confirmed** and the **secondary hypothesis was accepted**. These results are particularly surprising because the USDA's LILA measure is an indicator that includes both poverty rates and distance to supermarkets in the evaluation. One of the challenges with this indicator is likely in its binary nature: rather than being introduced on a scale, LILA denotes a region as either food secure or food insecure. Binary measures like this can create confusion because there are likely areas where food insecurity is prevalent but would not be considered food insecure because the majority of residents are food secure.

By contrast, poverty rate is measured on a scale from 0-100. This scale provides a gradient that can be used to better understand the relationship between poverty and diabetes as compared to the binary LILA measure. Clearly being able to see this relationship paints a strong image of what this relationship is.

Thus, while the results of the primary and secondary hypothesis defy intuition, they stand to reason when you consider the way the statistical methods would go about interpreting the results of the underlying data.

Conclusions
Based on the results of this study, the **primary hypothesis results are not conclusive**. While the data communicates that a relationship between diabetes and food access does not exist, the fact that a measure of food access – poverty rates – strongly correlates with diabetes prevalence calls these results into question. It is likely that the binary nature of this measure is the reason for the inconclusive and confusing results. To make this data conclusive, the underlying methodologies used by the USDA would need to be investigated and likely reverse engineered to be built on a scale – for example from 1-10 where 1 is least food secure and 10 is most food secure.

By contrast, the **secondary hypothesis results are definitively conclusive**. By every measure, the null hypothesis was rejected and a strong relationship was identified. Additionally, this relationship was confirmed by literature outside of the analysis from this study.

The ultimate conclusion that can be drawn from this study is the importance of food affordability to reduce the prevalence of diabetes. The data suggests that individuals who are more capable of making healthy purchase options do so. Thus, were this study to be used for future policy decision making food affordability should be considered at the forefront.

Replication
Were this study to be replicated or revised in the future, it would be helpful to have data on diabetes rates at a Census Block level. This would allow for analysis not only at a county level and provide the opportunity for greater insight into the relationship between food access and diabetes.

Addendum

All data analysis and processing is available for viewing on GitHub. **I completed this project independently.**

The Pandas, Scipy, MatPlotLib, and StatsModels libraries were used in the development of this project.

Sources

"How Many Americans Have Diabetes?" USA Facts. [Online]. Available: https://usafacts.org/articles/how-many-americans-have-diabetes/. [Accessed between: Oct. 31 - Nov. 29, 2024].

"Data & Research: Diabetes Public Health," Centers for Disease Control and Prevention. [Online]. Available: https://www.cdc.gov/diabetes/php/data-research/index.html.

"U.S. Diabetes Surveillance System," Centers for Disease Control and Prevention. [Online]. Available: https://gis.cdc.gov/grasp/diabetes/diabetesatlas.html.

"Food Access Research Atlas," U.S. Department of Agriculture, Economic Research Service. [Online]. Available: https://www.ers.usda.gov/data-products/food-access-research-atlas/.