OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Economic Conditions, Supermarket Distance, and Modeled Diabetes Prevalence Across U.S. Counties

Rohan Tiwari
tiwarirohann@gmail.com

ABSTRACT

Diabetes is one of the most common and expensive chronic diseases in the United States, and the burden is not evenly distributed geographically. The issue of food accessibility in policy discourse often revolves around the concept of distance, namely how far people live from supermarkets. This paper poses a narrower, descriptive question: in the context of U.S. counties, does the modeled prevalence of adult diabetes correlate more with economic factors or with a distance-based measure of poor supermarket accessibility? In an ecological design using publicly available county data from the USDA Food Environment Atlas, we related the modeled 2019 diabetes rate from the CDC to the poverty rate, median household income, food insecurity, and the percentage of the population living far away from a supermarket using correlations, multiple regressions with state-clustered inference, comparison within states, and out-of-sample predictions. There are two limitations inherent in this data that influence our findings. First, methods documentation for these CDC estimates indicates that county diabetes prevalence was estimated with a model that uses county poverty as an input variable, thus, some of the relationship between the latter and the outcome is built into the model; second, the economic variables (from 2021 to 2023) were measured after the outcome (in 2019). Within these constraints, poverty showed the strongest correlation with the modeled diabetes rate ($r = +0.44$), while the distance measure demonstrated a weak, slightly negative correlation ($r = -0.08$). A comparison of the variables within states yielded similar results ($r = +0.33$ and $r = -0.01$, respectively), and the association between the poverty rate and diabetes was still significant after adjusting for clustering at the state level. Together, the four predictors provided marginally better fit compared to the poverty rate alone ($R^2 = 0.196$ versus 0.190) and made little difference in out-of-sample predictions; a simple state-average baseline predicted more accurately than the regression in 199 out of 200 random samples. State-level analysis and obesity rates (which are not modeled from county poverty rates) also placed the economic measures ahead of store distance. These are correlations with modeled estimates, not causal effects, and they involve one distance measure rather than the entire food desert concept, which includes low income.

Keywords: *social determinants of health; food access; diabetes; poverty; ecological study; small-area estimates; predictive validation*

1. INTRODUCTION

Chronic illnesses make up a substantial proportion of the expenditure incurred in health matters in the United States, and the United States invests more per person in health issues than its peers. Nonetheless,

cases of diabetes and obesity have continued to occur at high rates, and in some instances, their frequency is increasing. It is estimated that about 40 percent of adult Americans suffer from obesity (Emmerich et al., 2024), while about 38 million people are living with diabetes, amounting to about 11.6 percent of the American population; about 29.7 million of these have received a diagnosis (CDC, 2024).

These burdens come at a price. In 2022, it is estimated that the cost of diagnosed diabetes was $412.9 billion, which includes $306.6 billion in direct medical care expenses and $106.3 billion in decreased productivity and early death (Parker et al., 2024). Medical costs related to obesity are around $173 billion each year (Ward et al., 2021). There is an uneven distribution of this problem, thus it becomes a public health problem rather than a purely medical issue. These burdens mostly affect low-income areas, and the question is obvious: what is it about poverty-stricken areas that makes them unhealthy?

One solution is found in the social determinants of health and the built environment. The factors that influence health are not just behavioral; they include the environments in which we exist, such as our income level, access to nutritious food, and the distance we must travel to obtain food (Braveman et al., 2011). One particular interpretation of this approach has gained wide acceptance within the United States. Food desert theory posits that people in poor, low-access areas eat unhealthily due to lack of access to nutritious food, and has justified the expenditure of public funds in the creation of supermarket facilities in these areas.

But there is little evidence that greater proximity causes improved health outcomes. In the studies which see opening of a new supermarket as a natural experiment, the impact of such opening on dietary habits and BMI of the residents is insignificant, while their perception of access is positively changed (Cummins et al., 2014). It seems that access itself is multidimensional; it involves distance from a supermarket, availability of its products, willingness to shop there and quality of the products offered (USDA ERS, 2023). But the distance is just one dimension. That is why we should consider not a test of the entire food desert hypothesis but a more narrow comparison: is estimated diabetes prevalence more related to an indicator of access or economic well-being in counties?

Three descriptive questions will be answered using freely accessible, federal data on the county level. First, what is the degree of association between poverty, income, food insecurity, and a distance-based low access metric and county modeled diabetes rates, respectively? Second, how do these associations change when the four variables are used together, uncertainty is permitted to vary within states, and counties are compared to those in their own states alone? Third, can this set of variables be used to build a model that outperforms other baselines, including those in which entire states have been excluded from the estimation? Related literature is reviewed in Section 2. The data and construction of the dependent variable are described in Section 3. Methods are explained in Section 4. Results are reported in Section 5, Section 6 discusses the findings, and Section 7 concludes.

2. BACKGROUND AND RELATED WORK

There are three key terms that appear throughout the course of this paper. Social determinants of health refer to the economic and social conditions that influence health other than access to healthcare. Built environment refers to the physical context in which individuals live. It encompasses the presence of surrounding stores as well as the travel time to get access to fresh foods. Food desert can be defined as the combination of low-income status and low access to grocery stores (USDA definition of "low-income, low-access," U.S. Congressional Research Service, 2021). Due to this dual economic and geographic definition, distance alone cannot be used to determine food deserts.

This is the difference that sets the limits for this paper. The food desert theory as a whole and the health effects of introducing the supermarket are not tested here. We focus only on the element of that theory, which is the distance to the supermarket, and we use it against other indicators that measure economic resources separately. There is no variable used in the analysis that includes both of these factors simultaneously.

There have been several prior connections made between the food environment and health via diet. For instance, Larson, Story, and Nelson (2009), in their review of the literature on neighborhoods, diet and disease, discovered that poorer districts lack grocery stores and access to fresh produce, which is associated with poor diets and obesity. Furthermore, Seligman, Laraia, and Kushel (2010), analyzing national data, concluded that food insecurity was associated independently with chronic diseases, such as diabetes. Evidence from a natural experiment discussed in Section 1 is an additional reason why one should not expect that proximity improvements will automatically have positive impacts (Cummins et al., 2014). Multiple studies analyzed associations between the food environment and diabetes or heart disease using the USDA Food Environment Atlas data at the county level. However, there are few studies that compare at the national level economic and distance variables and their relationship to health outcomes while paying attention to construction of health outcome data.

3. DATA

3.1 Source, unit of analysis, and sample

All variables were selected from the USDA ERS Food Environment Atlas (USDA ERS, 2025), which is a national compilation of federal survey and administrative data. All the variables were collected from the map service layers offered by the Atlas site (Appendix A) using the five-digit county FIPS code called "geoid". The resulting dataset includes information on 3,143 U.S. counties. In each analysis, all counties with complete information on the corresponding variables were included. There is no 2019 diabetes estimate in the Atlas for New Jersey's 21 counties or two census areas in Alaska, while the information about income and poverty is missing in one Hawaii county (Kalawao), so there are 3,119 counties in the sample for diabetes regression. The map of counties (Figure 1) covers the 3,108 counties of the contiguous U.S. from the data set, where 3,087 counties have a diabetes estimate.

3.2 Variables

There are six variables that will be used. These include the adult diabetes prevalence rate (primary outcome), adult obesity prevalence rate (secondary outcome at the state level), and the four predictors, namely poverty rate, median household income, food insecurity, and lack of store access. The variables, together with their roles, years, geographical levels, and method to derive state-level value are shown in Table 1.

Variable	Role	Year	Level	State value used in Section 5.4
Adult diabetes rate (%; CDC modeled estimate)	Outcome	2019	County	Unweighted mean of county values
Adult obesity rate (%)	Secondary outcome	2022	State	Published state value
Poverty rate (%)	Predictor	2021	County	Unweighted mean of county values
Median household income ($)	Predictor	2021	County	Unweighted mean of county values
Low food access (% of population far from a supermarket)	Predictor	2019	County	Unweighted mean of county values
Food insecurity (% of households)	Predictor	2021–23	State	Published state value

Table 1. *Variables, roles, years, geographic levels, and state aggregation rules. Source: USDA ERS Food Environment Atlas; diabetes and obesity estimates originate from the CDC. State-level values are repeated across all counties of the state in the county analysis.*

3.3 How the diabetes outcome was constructed

Variable pct_diabetes_adults19 representing diabetes is not the number of cases that were diagnosed. Instead, it is a model-based estimate of small area and it is produced by the Centers for Disease Control and Prevention (CDC). The CDC uses the U.S. Diabetes Surveillance System, which estimates prevalence for each county by fitting a statistical model to Behavioral Risk Factor Surveillance System survey data and predicting prevalence for all counties, even those with little to no responses. There was a change in the procedure that is being used. Older CDC documentation describes a Bayesian multilevel model that uses age group, sex, and race/ethnicity as covariates (CDC, n.d.-a). Estimates released from 2016 onward use a single year of BRFSS data in a multilevel model that combines respondents' age, sex, and race/ethnicity along with poverty at the county level and county-and-state-level contextual effects (County Health Rankings & Roadmaps, 2026). The CDC's PLACES method, used for later county estimates, likewise includes the county-level percentage of adults below 150 percent of the federal poverty level as a covariate (CDC, n.d.-b). Since the 2019 series belongs to the latter period and both methods include county poverty, we assume that county poverty is a covariate of the outcome.

There are three implications for this with regards to analysis. First, the relationship between poverty and diabetes is an imperfect measure of the relationship in question. A model that predicts diabetes based on county-level poverty yields estimates that increase with poverty levels in cases where the number of

responses in the surveys used is low, and this is especially prevalent in small counties where the model relies more heavily on the covariates used. Therefore, the correlation with poverty observed below is mechanical to some extent. Second, the overlap is partial rather than total: the 2021 Atlas poverty rate used here is not the same series as the poverty input to the CDC model. Finally, the model may also reduce associations with those variables that are not included in the list of the variables that the model takes as input, such as distance to the store.

An imperfect check of circularity may be performed using current data. State obesity prevalence (Section 5.4) represents state-level survey estimates and not small area county model estimates with poverty as an input, so there is no circularity in this instance in the same sense. A more precise check involves performing the county level comparison using a measure of diabetes that does not incorporate poverty as an input. This check is not included in this report and is the first priority for further research (Section 6.4). The within state comparison of Section 5.5 eliminates the state-level component of the CDC model, but not the county poverty input component.

3.4 Timing of the variables

These variables are not from one particular year. While the dependent variable and the access to stores variable are from 2019, poverty and income variables are from 2021, food insecurity from 2021 to 2023 and obesity from 2022 (Table 1). It is thus not one cross-sectional analysis but rather comparison within a time frame of about four years, which include the period of the COVID-19 pandemic. Economic predictors being collected after the outcome variable, they cannot be interpreted as preceding conditions. This comparison is based on an assumption that ranking of counties according to their economic development changes slowly enough to allow using of 2021 values instead of 2019 ones. This assumption is least reliable in the case of strong influence of economic shocks and responses during the pandemic period on counties' economies and also because of food insecurity measured in 2021-2023, reflecting the subsequent increase of food prices. Rerunning of the analysis with the use of poverty and income estimates in 2019 from the Small Area Income and Poverty Estimates by the Census Bureau (SAIPE; U.S. Census Bureau, n.d.) would validate this assumption.

3.5 What the low-access indicator measures

The "low food access" measure considers distance alone. It is the share of a county's population that lives over one mile from the closest grocery store in urban regions, or over ten miles in rural regions. This measure ignores food costs, car ownership among households, public transportation, and food quality, and it also does not include the low-income requirement from the USDA definition of a food desert. As a measure applied to all of a county's population, the "low food access" measure could also be obscuring concentrations of low food access within counties, as well as making an equal percentage represent entirely different situations in a densely-populated city versus a rural county. Results for this measure should be interpreted as such.

3.6 Geographic levels and aggregation

Obesity and food insecurity indicators are collected only at the state level. For the county-level analysis, food insecurity indicators from each state were replicated for all of the counties within that state. The replicated values are not independent observations at the county level and are unable to capture any variability within a state; hence, the interpretation of county-level results on food insecurity should be made cautiously, and inference is clustered by state (Section 4.3).

The state-level variables are calculated as the unweighted averages of the variables' values for the state's counties. The unweighted averages are calculated this way in order to preserve the link between the analysis at the two levels, because each county is an observation of local conditions, but this approach assigns equal weights to the counties with very different populations. An unweighted average of county rates is not a population prevalence rate for the state, while population-weighted averages would more accurately represent the rates at the state level, and different weighting schemes could influence the results, so the state-level results show average conditions in the states' counties, but not population prevalence rates.

3.7 Data preparation

Data preparation was done using Python. The indicators were obtained from the service layers of Atlas mentioned in Appendix A and then joined using the five-digit FIPS code. All records with missing or unrealistic values were deleted from the dataset. There was no imputation of values or any sort of outliers' handling done. Every analysis is done using all complete records for their particular variables; that is why there are different numbers of observations in each analysis (Appendix A).

4. METHODS

4.1 Design and models

The study design is both an observational study and an ecological study; it compares counties on the basis of measurements between 2019 and 2023, not changes over time. It can describe relationships but it cannot demonstrate causality. Ordinary least squares (OLS) regressions were run for the rate of diabetes. Simple regression models analyze each independent variable in isolation, while the multivariate regression model uses all four predictors, and each of the coefficients in this model reflects the relationship between that predictor and diabetes, controlling for the remaining three. All models have an intercept term. The predictors were converted into z-scores and thus have comparable coefficients despite being measured in different units.

In order to estimate only the effects on counties compared to other counties in the same state, we also run regressions with state fixed effects, meaning an intercept for each state. This will capture all the information that is the same in each state, including food insecurity in the state, policies in the state, regional differences, and the state-specific component of the CDC outcome equation. This does not control for differences across counties within the same state.

4.2 Measuring association

Two correlations are provided in this report. The first one, Pearson's r, represents how strongly two variables are linearly related, and for simple regression, it is equal to the sign of the square root of R^2, the proportion of variance in the dependent variable explained by an independent one. The second correlation, Spearman's ρ, is based on ranks. The confidence interval for the correlation coefficient is based on the Fisher's z transformation. Due to the large number of counties, even small relationships will be statistically significant, therefore, our attention is drawn mainly to effect size (r and R^2) rather than to statistical significance alone (Wasserstein & Lazar, 2016). R^2 is presented in percentage form, the regression coefficients and error predictions are in percentage points of diabetes rate, and the difference between two percentages is measured in percentage points.

4.3 Inference when observations are not independent

The traditional p-values and confidence intervals rely on the assumption that the counties are independent observations. This assumption does not hold here, as the counties in the same state share a common food insecurity level, state laws, and the effects of the state-level in the CDC outcome model, and neighboring counties are similar to each other. Standard errors clustered by state (Cameron & Miller, 2015) are therefore also provided; such standard errors permit errors to be correlated among all counties in the same state. As there are only 50 clusters, they are approximate; moreover, they ignore the correlation between the neighboring counties from different states. Moran's I index (Moran, 1950) of the regression residuals is computed directly to quantify the spatial dependence. The counties are neighbors if they share a boundary line (using the 2010 Census boundaries that form the basis for Figure 1, there are 3,111 counties that have at least one neighbor). Significance is measured via 999 random permutations. For the state-level analysis, each state represents one observation, and confidence intervals are also provided.

4.4 Prediction tests

Explaining variations is one type of test, but predicting new observations is another type that is even more difficult. For the first prediction test, we randomly divided the counties into a training set (80%) and a test set (20%), using a predetermined random seed number. Means and standard deviations of predictor variables were estimated for the training set and used to normalize values of the test set. The regression model was tested against two baselines – the national-mean baseline predicts each test county with the average rate of the training counties, while the state-mean baseline predicts each test county with the average rate of the training counties in the same state. To illustrate the contribution of the other three factors besides poverty, we fitted the regression model based on only the poverty factor. To avoid the problem that a single split may not be representative enough, we did the same random split 200 times with different seeds. The evaluation measure is root mean squared error (RMSE) in percentage points.

Random partitioning results in more optimistic performance than it should, as the adjacent counties tend to be similar. In the second experiment, the whole state is omitted with five-fold cross-validation grouped by state, such that the model predicts the states that it hasn't seen before. It is not possible to calculate the

state-mean baseline in this setting, since the omitted state does not have any counties in the training set, and hence, employing its average would require test information for prediction.

4.5 Model checks

The measure of multicollinearity is done using the Variance Inflation Factors (VIFs). The VIFs are medium (poverty 2.6, income 2.5, food insecurity 1.2, low access 1.0): below the normal level of 5, but sufficiently large for instability of the poverty and income coefficient estimates. There are also problems of non-normal and heteroscedastic residuals.

The previous version of this analysis showed a Durbin-Watson value of roughly 1.60. This statistic is an indicator of the correlation among sequential rows in the data, and hence, depends on the order in which the rows are arranged. Since the counties have been sorted according to their FIPS numbers, this is an indication of the similarity between counties that are adjacent to each other in terms of their FIPS number, and thus, are likely to belong to the same state. It is therefore not a test of geographic correlation, and we no longer rely on it; spatial dependence is assessed with Moran's I instead (Sections 4.3 and 5.5).

4.6 Software and reproducibility

Analysis was conducted in Python 3.12 utilizing the packages pandas, NumPy, SciPy, statsmodels, scikit-learn, Matplotlib, and GeoPandas (Appendix B). Code and processed data will be shared in an open access database prior to publishing (see Data Availability).

5. RESULTS

5.1 The distribution of diabetes prevalence

In Figure 1, county-level modeled diabetes prevalence is shown. Counties with high prevalence do not occur randomly but cluster. The deepest color occurs in the Deep South and Appalachia region that consists of Mississippi, Alabama, Georgia, and Louisiana; West Virginia, east Kentucky, and east Tennessee: the diabetes belt (Barker et al., 2011).

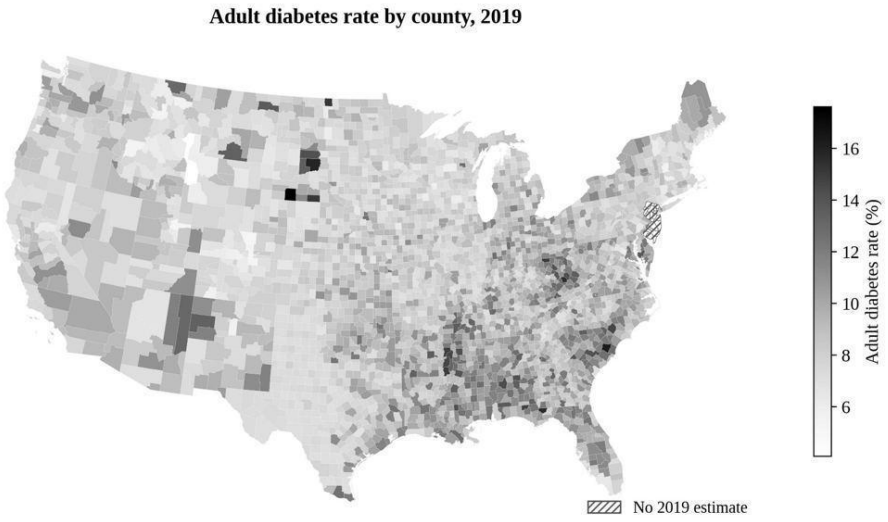


Figure 1. *Diabetes Prevalence Among Adults by County, 2019 (contiguous U.S.; 3,087 counties for which estimates are available). The darker shading means a higher prevalence. For New Jersey, there is no 2019 data available in the Atlas for its 21 counties, which are hatched. It is a modeled CDC estimation. Data source: USDA ERS Food Environment Atlas.*

The diabetes belt generally coincides with the zone of concentrated and enduring poverty of the South and Appalachia. It does not coincide with distance to the supermarket, since the locations with the greatest distances from the supermarkets are located in the lightly populated Mountain West and part of the Great Plains, most of which have lower modeled rates of diabetes. There are two things to bear in mind. First, the outcome model includes county-level poverty as one of its inputs (Section 3.3), which means that such a coincidence of patterns would be expected a priori. Second, these areas vary in age distribution, racial composition, and rural character.

5.2 One predictor at a time

Table 2 shows the correlation of the variables one by one. Poverty has the highest correlation with the modelled rate of diabetes (r = +0.44), and explains 19.0% of the variation in the modelled rate. Median income is ranked second and has a negative relationship (wealthier counties have lower rates of diabetes). The third is food insecurity, which has a very low correlation, while low access is essentially unrelated to it.

Predictor	Pearson r	Spearman ρ	95% CI (r)	p	R^2 (%)	p (state-clustered)
Poverty rate	+0.44	+0.38	[+0.41, +0.46]	<0.001	19.0	<0.001
Median income	−0.31	−0.29	[−0.34, −0.28]	<0.001	9.6	<0.001
Food insecurity	+0.18	+0.18	[+0.14, +0.21]	<0.001	3.1	0.18
Low food access	−0.08	−0.03	[−0.11, −0.04]	<0.001	0.6	0.005

Table 2. *Association of each predictor with the modeled county diabetes rate (n = 3,119). The conventional confidence intervals and p-values assume independent counties and are too narrow; the last column gives p-values from*

single-predictor regressions with standard errors clustered by state (Section 4.3). With clustering, the food insecurity association, a state value repeated across counties, is no longer statistically significant. The poverty association is partly built into the outcome (Section 3.3).

Here are two important points. One, statistical significance does not equate to practical significance. The low access variable has a Pearson's correlation value of −0.08, Spearman's correlation value close to zero (ρ=−0.03), and accounts for less than one percent of the total variance. Moreover, the negative sign of the low access variable is contrary to expectations that increasing distances increase the prevalence of diabetes, and it is small enough to have no substantive significance in either direction. Two, the variables representing the economic factors have similar values in Pearson's and Spearman's correlation, meaning that their relationships are not based on outliers. Figure 2 depicts all the four correlations.

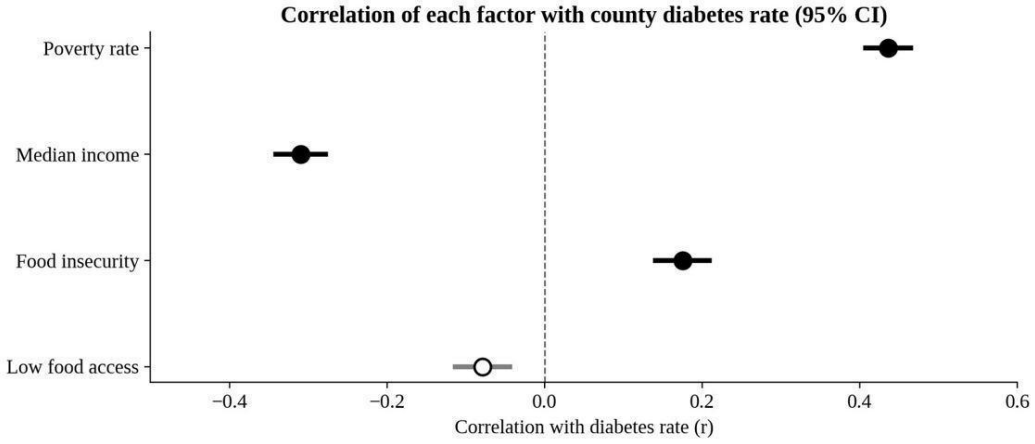
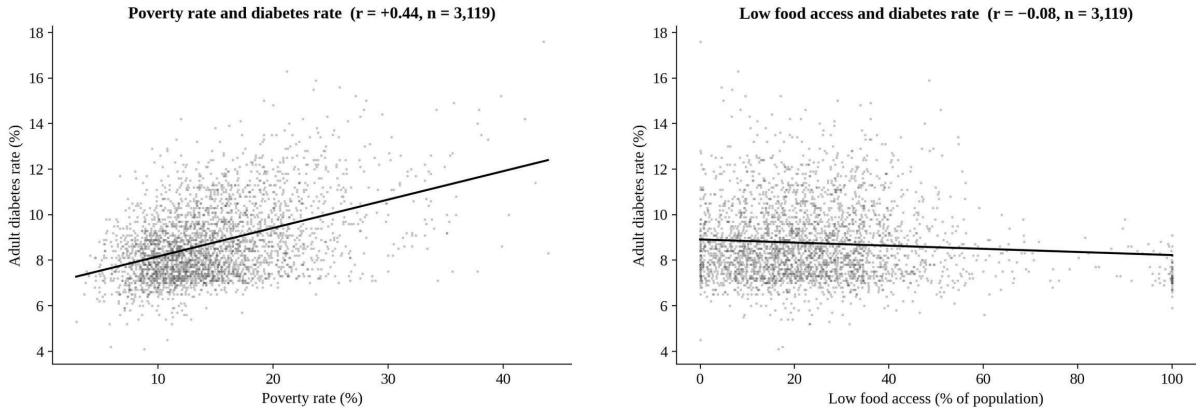


Figure 2. *Correlation of each predictor with the modeled county diabetes rate with 95% confidence intervals assuming independence between counties. The filled circles indicate those with correlations where $|r| \geq 0.15$ while the open circle is near zero.*

Figure 3 is the plot of poverty versus modeled diabetes and shows an unmistakable increasing trend for all 3,119 counties. In Figure 4, there is a difference with the plot of the low-access variable versus modeled diabetes.



Figures 3 and 4. *Left: County poverty rate and modelled diabetes rate correlation (r = +0.44). Right: County low-access indicator and modelled diabetes rate correlation (r = −0.08). The access indicator is nearly free from any relationship to modelled diabetes.*

5.3 The combined model and its predictive performance

Combining all four predictors in one regression refines but does not refute the one-factor perspective. They collectively account for 19.6 percent of the variance of the modeled diabetes prevalence (R^2 = 0.196), which is merely 0.6 percentage points more than poverty on its own (19.0 percent). This supports a descriptive ordering of the four indicators. It does not imply, however, that poverty plays the primary role in driving diabetes because some portion of the poverty effect is already embedded in the dependent variable, among other factors. Table 3 presents coefficients for all four indicators.

Predictor	Coef. (pp per SD)	95% CI (state-clustered)	p (conventional)	p (state-clustered)	VIF
Poverty rate	+0.777	[+0.59, +0.96]	<0.001	<0.001	2.59
Median income	+0.117	[−0.03, +0.26]	0.004	0.12	2.50
Food insecurity	+0.041	[−0.27, +0.35]	0.14	0.79	1.15
Low food access	−0.089	[−0.18, +0.00]	<0.001	0.06	1.00

Table 3. *Multivariable model for the modeled diabetes rate (R^2 = 0.196, n = 3,119). Predictors are standardized, so each coefficient is the change in prevalence (percentage points) associated with a one-standard-deviation increase, holding the others constant. Conventional p-values assume independent counties; state-clustered intervals and p-values allow errors to be correlated within states (Section 4.3). With clustering, only the poverty coefficient is statistically significant.*

The most surprising coefficient is income, which becomes positive after introducing poverty into the regression, despite being negative on its own (Table 2). This can be explained by the hypothesis that keeping the poverty rate constant but increasing median income results in an increase in inequality, and therefore the coefficient can partially be attributed to this. Alternatively, it can be attributed to the splitting of two correlated indicators of the same income distribution. The coefficient is statistically insignificant when clustered by states (p = 0.12). For these reasons, we do not interpret the multivariable income coefficient and base the ranking on the single-predictor results.

Summary statistics of the prediction tests are presented in Table 4. Results are rather inconsistent. Most important comparisons are made with respect to the state-mean baseline. In terms of the original 80/20 split, assigning each county the state mean of other counties (RMSE = 1.32 pp) was a more accurate predictor of modeled diabetes compared to the four-factor regression (1.43 pp). Importantly, the state-mean baseline performed better in the majority of random splits (199 out of 200 random splits; RMSE 1.36 pp versus 1.46 pp). That is, knowledge about which state each county belonged to helped to predict modeled diabetes more effectively than knowing the values of poverty, income, food insecurity, and store distance for that county. The four-factor regression still proved superior to the national-mean baseline (RMSE 1.57 pp in terms of the original split and 1.62 pp on average), and there were no clear signs of overfitting: original R^2 (training) and test R^2 values (0.201 and 0.172 respectively) are quite reasonable. The three factors aside from poverty contribute almost nothing out-of-sample: across 200 random splits, a poverty-only regression had a mean RMSE of 1.460 pp versus 1.456 pp for the four-factor model.

Even when full states were withheld, the four-factor regression outperformed the national mean (average RMSE 1.50 compared to 1.63 percentage points; average test $R^2 = 0.11$) but was inconsistent across folds (Table 4b). In the fold that held out Texas, Ohio, West Virginia, and seven other states, the R^2 was negative (-0.06), indicating that it predicted those states worse than even the training mean. The poverty regression performed better: it had a positive R^2 in all folds and an out-of-fold R^2 of 0.18, whereas the four-factor regression had a value of only 0.13. This is further evidence that the extra factors made the predictions less reliable, rather than more accurate.

Prediction test	National mean RMSE	State mean RMSE	Four-factor RMSE	Four-factor R^2	Poverty-only RMSE	Poverty-only R^2
Random 80/20 split (seed 42)	1.566	1.323	1.425	0.172	1.431	0.165
200 random splits: mean (SD)	1.622 (0.047)	1.363 (0.049)	1.456 (0.041)	0.192 (0.032)	1.460 (0.041)	0.188 (0.031)
Hold out whole states: mean of 5 folds (SD)	1.632	—[a]	1.500 (0.180)	0.109 (0.111)	1.462 (0.128)	0.153 (0.048)
Hold out whole states: pooled out-of-fold	1.635	—[a]	1.508	0.133	1.466	0.181

Table 4. *Prediction accuracy for the modeled county diabetes rate (RMSE in percentage points; lower is better; R^2 on held-out counties). The state-mean baseline beat the four-factor regression in 199 of the 200 random splits. [a] Cannot be computed by design: a held-out state has no training counties, and using its own mean would leak the prediction targets.*

Fold	States held out	n	National mean RMSE	4-factor RMSE	4-factor R^2	Poverty RMSE	Poverty R^2
1	Alaska, Arkansas, District of Columbia, Hawaii, Maine, Ohio, Oregon, South Dakota, Texas, West Virginia	623	1.721	1.775	−0.064	1.620	0.113

Fold	States held out	n	National mean RMSE	4-factor RMSE	4-factor R^2	Poverty RMSE	Poverty R^2
2	Colorado, Georgia, Iowa, Maryland, Minnesota, Montana, Nevada, New Hampshire, Washington, Wisconsin	627	1.559	1.372	0.084	1.353	0.109
3	Alabama, California, Connecticut, Indiana, Massachusetts, North Carolina, North Dakota, Oklahoma, Virginia, Wyoming	625	1.564	1.442	0.134	1.425	0.155
4	Arizona, Illinois, Kentucky, Michigan, Nebraska, New York, Pennsylvania, Rhode Island, South Carolina, Utah	622	1.526	1.333	0.235	1.340	0.228
5	Delaware, Florida, Idaho, Kansas, Louisiana, Mississippi, Missouri, New Mexico, Tennessee, Vermont	622	1.788	1.577	0.153	1.571	0.160

Table 4b. *Fold-by-fold results of the held-out-states test (five folds grouped by state; scikit-learn GroupKFold, version 1.8). "Poverty" is the poverty-only model. Negative R^2 means the model predicted the held-out states less accurately than the mean of the training counties.*

5.4 A state-level analysis

Since obesity and food insecurity measures are provided only for states, and since using state values at the county level would produce too narrow an interval range at the county level, we ran a separate analysis with one value per state: 49 states plus DC (NJ is not included since there are no diabetes estimates for 2019 for NJ).

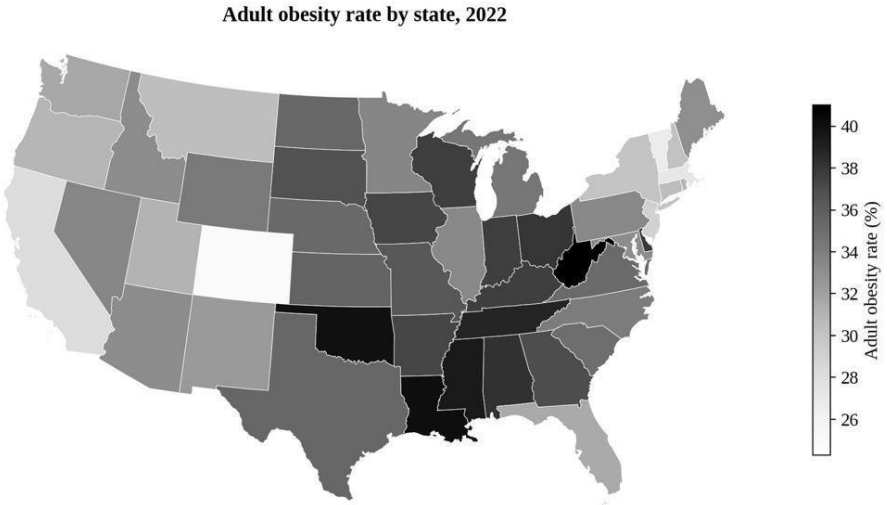


Figure 5. *Obesity in Adults Across States in 2022 (contiguous US and Washington DC; Alaska - 32.1% and Hawaii - 25.9% not shown). The darker the shade, the greater the prevalence. Source: USDA ERS Food Environment Atlas (field pct_obese_adults22); data from the CDC.*

Obesity map resembles that of diabetes map as well. States where the proportion of obese adults exceeds 38 percent are West Virginia, Louisiana, Oklahoma, Mississippi, Tennessee, Alabama, and Ohio, whereas those states with the lowest rates of obesity are District of Columbia, Colorado, Hawaii, Vermont, and Massachusetts. Mountain West states with high number of people who lack access to supermarkets remain under the national median of 33.6 percent (with Wyoming being an exception since its level of obesity amounts to 34.3 percent). In order to conduct the analysis, predictors at the county level and diabetes prevalence were averaged by states without weights (Section 3.6).

Predictor	Diabetes r [95% CI]	p	R^2 (%)	Obesity r [95% CI]	p	R^2 (%)
Poverty rate	+0.66 [+0.47, +0.79]	<0.001	43.3	+0.43 [+0.17, +0.63]	0.002	18.8
Median income	−0.44 [−0.64, −0.18]	0.001	19.5	−0.71 [−0.83, −0.54]	<0.001	50.1
Food insecurity	+0.44 [+0.18, +0.64]	0.001	19.3	+0.55 [+0.32, +0.72]	<0.001	29.9
Low food access	−0.20 [−0.45, +0.08]	0.159	4.1	+0.01 [−0.27, +0.29]	0.945	0.0

Table 5. *State-level analysis (n = 50: 49 states and DC). County variables are unweighted state means of counties; obesity and food insecurity are published state values. Confidence intervals are Fisher-z intervals computed from the reported r. With 50 observations, intervals are wide, and the ordering among the three economic measures is not precisely estimated.*

The correlations at the state level are more evident (Table 5). Poverty accounts for 43 percent of the variation in state modeled diabetes prevalence (as against 19 percent for county-level variations) while median income accounts for 50 percent of the variation in state obesity prevalence. The variable indicating low access is not statistically significant in either case. The intervals are wide, however: the data are consistent with a poverty–diabetes correlation anywhere from about 0.47 to 0.79, and the three economic measures cannot be confidently ranked against one another. What the state analysis does support is the broader pattern that economic measures are more strongly associated with both outcomes than store distance is.

In another way, the results regarding obesity provide useful insights. State obesity estimates do not involve circularity because these estimates are based on surveys and not on an estimation of a county level model that includes poverty in its specification (as explained in section 3.3). In terms of obesity, economic indicators ($r = -0.71$ for median income; $r = +0.43$ for poverty) still clearly outperform the indicator that reflects store distance ($r = +0.01$).

A few considerations could be behind the stronger relationships found following the aggregation. Averaging removes random error in the county-based estimates, which would otherwise weaken correlations. Food insecurity is a state-level indicator in the county-level estimates and cannot account for variations within states, which likely accounts for most of the rise in its explained variance from 3% to 19%. The average also reduces the variability not accounted for in the model's four predictor variables, yielding mostly differences in broader geographic regions, which correlate with poverty. Also, some effects, including Medicaid policy, act on a state-wide level. Lastly, correlations between aggregate units tend to be greater than the correlation of their constituent parts (Robinson, 1950; Openshaw, 1984). In this case, the state diabetes indicators are simply averages of the output of county models using poverty as an input.

5.5 Robustness checks

Inference allowing for state-clustered errors. Allowing for clustering of the errors within states increases uncertainty substantially, but the association with poverty is still significantly positive, both by itself (Table 2) and in the multivariable regression, in which it has a coefficient of 0.78 percentage points per standard deviation (95% CI +0.59 to +0.96; Table 3). In the multivariable regression, the low-access coefficient is no longer statistically significant ($p = 0.06$), nor are the income and food insecurity coefficients.

Spatial Autocorrelation. The residual for the four-factor model exhibits a pronounced degree of spatial clustering (Moran's $I = 0.34$, permutation $p = 0.001$; Moran's I for the modeled diabetes rate itself is 0.43). The counties for which the model predicts too high or too low diabetes rates are clustered together geographically. This demonstrates that the precision provided by the county p-values is an exaggeration, and also suggests that regional factors outside of the model account for much of the remaining variance.

Comparison within the same state. Table 6 presents the comparison of the counties solely with those from the same state. State-level fixed effects account for 31.7 percent of the variance of the modeled diabetes prevalence rate, which is higher than the 19.6 percent accounted for by the four predictors combined. Within states, the poverty rate is positively correlated with the modeled diabetes rate (within-state $r = +0.33$; coefficient +0.53, $p < 0.001$), while the low-access variable is not ($r = −0.01$; coefficient −0.02, $p = 0.65$). Adding poverty to the state effects raises R^2 from 0.317 to 0.389; adding low access leaves it unchanged. Since the state effects capture state policy, regional variations, food insecurity, and the state level component of the CDC outcome model, this is the most convincing example of why the difference between poverty and distance to stores is not due to geographical differences between states. It does not eliminate the county level poverty variable from the outcome model or geographical variations of age, race and ethnicity, or rurality within the same state.

Model	Poverty	Median income	Low food access	R^2
Four-factor model, no state effects (Table 3)	+0.78 [+0.59, +0.96]	+0.12 [−0.03, +0.26]	−0.09 [−0.18, +0.00]	0.196
State fixed effects only	—	—	—	0.317
State fixed effects + poverty	+0.53 [+0.36, +0.70]	—	—	0.389
State fixed effects + low access	—	—	−0.02 [−0.10, +0.06]	0.317
State fixed effects + poverty, income, low access	+0.64 [+0.41, +0.86]	+0.14 [−0.06, +0.33]	+0.00 [−0.08, +0.09]	0.391
Within-state correlation with modeled diabetes (r)	+0.33	−0.21	−0.01	

Table 6. *Within-state comparison (n = 3,119 counties in 50 states). Coefficients are percentage points of modeled diabetes per standard deviation of the predictor, with 95% confidence intervals from state-clustered standard errors. State fixed effects give each state its own intercept, so food insecurity, a state-level value, cannot be estimated. Within-state correlations use deviations from each state's mean.*

Other checks have yet to be performed. The comparison has not yet been repeated using an indicator of diabetes that excludes the poverty variable, with measures for the economy from 2019, with control

variables at the county level for age composition, race and ethnicity, rurality, physical inactivity, and access to health care, or with state averages weighted by population. This list is provided as future research in Section 6.4.

6. DISCUSSION

6.1 What the analysis shows

Regardless of the different methods used, the economic factors seemed more related to the modeled diabetes prevalence than the low access measure based on distance. From the map, the distribution of diabetes and obesity was more closely associated with poverty than with the distance of stores; from the correlation, poverty and income showed a higher rank than the food insecurity and low access; from the state level analysis, the economic factors again ranked ahead of store distance for both diabetes and obesity; from the within-state comparison, poverty was linked to modeled diabetes but distance was not; and from the prediction tests, the four factors carried some signal that transferred to unseen states, although a county's state alone predicted better.

These findings should be stated narrowly. They show that, at county scale, this particular distance indicator carries little information about modeled diabetes prevalence. They do not test the food desert hypothesis as a whole, which concerns places that are both low-income and low-access, and they say nothing about the health effect of improving access for a given community. Nor do they prove that poverty is the dominant factor for diabetes development. The output variable is based on county poverty levels in its calculations, hence, part of the correlation can be attributed to this fact, as well as the fact that the four variables explain just one fifth of the variance of county differences, which is less than the state alone, and there is no control for other possible explanatory variables like age structure and rurality at the county level.

6.2 What it means for policy

In case distance were the key limiting factor, the solution would be geographic; open grocery stores, subsidize supermarkets, and use shuttles. These results do not indicate that the concept of proximity is of no significance but rather that at the county level the distance variable has less signal relative to economic variables. This result is in accordance with findings from natural experiments showing that opening grocery stores can lead to improved perceptions of access to healthy foods despite diet and weight remaining the same (Cummins et al., 2014). A reasonable implication is that evaluations of such food access interventions should take into account affordability and economic resources along with proximity and the need to examine combinations of all of these factors.

6.3 Limitations

- **Modeled outcome.** The diabetes rate is based on the small area estimates provided by the CDC, and there is a documented link between poverty in the county and the modeling inputs used to predict the

outcome in this time frame (Section 3.3). In other words, the poverty association is partly built into the outcome.

- **Timing.** The economic indicators were collected two to four years after the dependent variable. These act as proxies for the state of affairs in 2019, but have not been tested on 2019 data (Section 3.4).
- **Distance Indicator Only.** The low-access variable measures distance but not price, transportation, or quality, and when averaged over the entire county, it could hide local areas with poor access (Section 3.5).
- **Correlation, but not causation; Ecological Inference.** Findings relate to the levels of counties and states, and not to the individual level.
- **Omitted variables.** "Most powerful among the measures" does not mean "the most important variables overall." The fixed effects for each state control for variations across states, but variables related to the age structure, race and ethnicity, rural status, physical inactivity, and access to healthcare were omitted and may explain some of the relationship.
- **Correlation in observations.** The observations are spatially correlated (Moran's I of residuals = 0.34). Cluster-robust errors by state will capture correlations within states but not across state lines.
- **State Weighting.** State values for county-level variables are unweighted averages, which characterize the typical county rather than the state's prevalence.
- **Prediction.** State-level average performed better in 199 of 200 random iterations, and the four variables were less portable to unseen states compared to just the poverty rate.

6.4 Future work

A number of tests could be done to determine the strength of this comparison. The first is repeating the test using an outcome measure of diabetes that is not estimated based on the relationship between county poverty and this disease. This can be the prevalence of claims-based diabetes among Medicare beneficiaries or the mortality rate of diabetes as determined from death certificates. In this way, no part of the association is built into the outcome measure. Secondly, using 2019 poverty and income estimates from SAIPE to match the outcome year. Thirdly, including covariates such as age distribution, race/ethnicity, rural areas, physical inactivity, and insurance status, and testing whether poverty continues to have a stronger association with diabetes than distance to store within states.

Apart from the tests above, a more direct test of the food desert hypothesis would be possible using indicators like the number of people living in economically distressed households using SNAP or the Gini index measuring income inequality, as well as the USDA composite measure for low income and access at the census-tract level. Lastly, the findings from the test can be transformed to a format useful to a local partner in the form of a county report useful to a food bank or clinic.

7. CONCLUSION

With regard to the ecological assessment of the USDA's metrics of supermarket accessibility versus county health model metrics of diabetes prevalence, there was a clearly stronger association with the economic factors of poverty and income, as compared to a distance indicator of low supermarket accessibility that essentially lacked signal. This relationship was the same both when comparing the

counties in the context of within-state comparisons and state clustered inference and when looking at states overall, as well as the same relationship being present with obesity, an outcome not modeled from county-level poverty. Because the diabetes prevalence is modeled partly from county poverty, and the economic predictors were measured after the outcome, such results are hypothesis generating in nature.

DATA AVAILABILITY

These data are readily accessible via the Food Environment Atlas of the USDA Economic Research Service (https://www.ers.usda.gov/data-products/food-environment-atlas); the data related to the diabetes and obesity statistics have been derived from the CDC. In Appendix A, the service layer and source fields of each data set are provided. The analysis was conducted using a Python script which provides all analytical outputs; these scripts, together with the software used, will be made publicly available prior to publication.

REFERENCES

Barker, L. E., Kirtland, K. A., Gregg, E. W., Geiss, L. S., & Thompson, T. J. (2011). Geographic distribution of diagnosed diabetes in the U.S.: A diabetes belt. *American Journal of Preventive Medicine, 40*(4), 434–439. https://doi.org/10.1016/j.amepre.2010.12.019

Braveman, P. A., Egerter, S. A., & Williams, D. R. (2011). The social determinants of health: Coming of age. *Annual Review of Public Health, 32*, 381–398. https://doi.org/10.1146/annurev-publhealth-031210-101218

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources, 50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Centers for Disease Control and Prevention. (2024). *National diabetes statistics report*. U.S. Department of Health and Human Services. https://www.cdc.gov/diabetes/php/data-research/index.html

Centers for Disease Control and Prevention. (n.d.-a). *Methods to obtain county-level estimates and ranks*. U.S. Diabetes Surveillance System documentation. https://www.cdc.gov/diabetes/pdfs/data/calculating-methods-references-county-level-estimates-ranks.pdf

Centers for Disease Control and Prevention. (n.d.-b). *PLACES: Methodology*. https://www.cdc.gov/places/methodology/index.html

County Health Rankings & Roadmaps. (2026). *Diabetes prevalence* [Measure methods]. University of Wisconsin Population Health Institute. https://www.countyhealthrankings.org/health-data/population-health-and-well-being/quality-of-life/physical-health/diabetes-prevalence

Cummins, S., Flint, E., & Matthews, S. A. (2014). New neighborhood grocery store increased awareness of food access but did not alter dietary habits or obesity. *Health Affairs, 33*(2), 283–291. https://doi.org/10.1377/hlthaff.2013.0512

Emmerich, S. D., Fryar, C. D., Stierman, B., & Ogden, C. L. (2024). *Obesity and severe obesity prevalence in adults: United States, August 2021–August 2023* (NCHS Data Brief No. 508). National Center for Health Statistics. https://doi.org/10.15620/cdc/159281

Larson, N. I., Story, M. T., & Nelson, M. C. (2009). Neighborhood environments: Disparities in access to healthy foods in the U.S. *American Journal of Preventive Medicine, 36*(1), 74–81. https://doi.org/10.1016/j.amepre.2008.09.025

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika, 37*(1/2), 17–23. https://doi.org/10.2307/2332142

Openshaw, S. (1984). *The modifiable areal unit problem* (Concepts and Techniques in Modern Geography No. 38). Geo Books.

Parker, E. D., Lin, J., Mahoney, T., Ume, N., Yang, G., Gabbay, R. A., ElSayed, N. A., & Bannuru, R. R. (2024). Economic costs of diabetes in the U.S. in 2022. *Diabetes Care, 47*(1), 26–43. https://doi.org/10.2337/dci23-0085

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*(3), 351–357. https://doi.org/10.2307/2087176

Seligman, H. K., Laraia, B. A., & Kushel, M. B. (2010). Food insecurity is associated with chronic disease among low-income NHANES participants. *Journal of Nutrition, 140*(2), 304–310. https://doi.org/10.3945/jn.109.112573

U.S. Census Bureau. (n.d.). *Small Area Income and Poverty Estimates (SAIPE) program*. https://www.census.gov/programs-surveys/saipe.html

U.S. Congressional Research Service. (2021). *Defining low-income, low-access food areas (food deserts)* (Report IF11841). https://crsreports.congress.gov/product/pdf/IF/IF11841

U.S. Department of Agriculture, Economic Research Service. (2023). *Food access: Research and data on access to affordable, healthy food*. https://www.ers.usda.gov/topics/food-choices-health/food-access

U.S. Department of Agriculture, Economic Research Service. (2025). *Food Environment Atlas*. https://www.ers.usda.gov/data-products/food-environment-atlas

Ward, Z. J., Bleich, S. N., Long, M. W., & Gortmaker, S. L. (2021). Association of body mass index with health care expenditures in the United States by age and sex. *PLOS ONE, 16*(3), e0247307. https://doi.org/10.1371/journal.pone.0247307

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

APPENDIX A. DATA DICTIONARY

Description of each analytic variable including the USDA source layer and field; the geographic level and year; what the variable represents; and the sample size of data. The sample sizes for each county-level variable, n, represent the number of counties having that variable along with the diabetes rate. There are 51 observations in state-level layers (representing 50 states and DC), out of which 50 are utilized (NJ omitted).

Field (source column)	Source layer	Level / Year	Role	n
diabetes_2019 (pct_diabetes_adults19)	Health25_health / 2	County / 2019	Outcome	3,119
obesity_2022_state (pct_obese_adults22)	Health25_health / 4	State / 2022	Secondary outcome	50 of 51
poverty_rate_2021 (povrate21)	Socioeconomic25_Income / 13	County / 2021	Predictor	3,119
median_income_2021 (medhhinc21)	Socioeconomic25_Income / 12	County / 2021	Predictor	3,119
low_access_pct_2019 (pct_laccess_pop19)	Access25_Overall / 5	County / 2019	Predictor	3,120
food_insecurity_2021_23_state (foodinsec_21_23)	Insecurity25_Insecurity / 2	State / 2021–23	Predictor	50 of 51

Table A1. *Data dictionary. All values are percentages except median income (dollars). Source: USDA ERS Food Environment Atlas.*

APPENDIX B. REPRODUCIBILITY DETAILS

The analysis was done in Python 3.12, using pandas 3.0, NumPy 2.4, SciPy 1.17, statsmodels 0.15, scikit-learn 1.8, Matplotlib 3.10, and GeoPandas 1.1. The code imports the merged dataset with counties; computes the Pearson and Spearman correlations with Fisher-z confidence intervals; fits the univariate and multivariable OLS regressions (intercept included) with standard errors in regular and clustered-by-state versions and with state fixed effects; runs the 80/20 split (random seed 42), 200 repetitions of the 80/20 split (random seeds 0–199), and the held-out-state testing (five folds grouped by state, GroupKFold) showing results for each fold and aggregated out-of-sample results; compares the models with national and state mean baselines; computes the Moran's I of the residual; and creates the maps and plots. The counties are ordered by five-digit FIPS code before any splits. The predictors are standardized based on training data only. For the held-out-state test, the GroupKFold method from scikit-learn with states as groups is used; because the allocation of states to the folds changes across different versions of the GroupKFold implementation, version 1.8 is needed to replicate Table 4b. The Moran's I computation uses the contiguity matrix derived from the boundary files specified below (counties are neighboring if they share a boundary), and the permutation test with 999 permutations. The county and state boundaries are from the U.S. Census Bureau cartographic boundary file (2010 edition, via Plotly GeoJSON). Three counties whose FIPS codes changed after 2010 were matched to their 2010 boundaries: Oglala Lakota County, South Dakota (46102, formerly Shannon County, 46113); Kusilvak Census Area, Alaska (02158, formerly Wade Hampton Census Area, 02270); and Bedford County, Virginia (51019, which absorbed Bedford city, 51515).