Neighborhood Disadvantage and Health Outcomes: A Census Tract-Level Analysis

1. Data & Decisions

This analysis draws on two publicly available datasets. The primary source for neighborhood-level socioeconomic characteristics is the American Community Survey (ACS) 5-year estimates (2015–2019), accessed programmatically via the tidycensus R package. Health outcome data come from the CDC PLACES 2021 release (census tract level), which was itself constructed using ACS 2015–2019 estimates as a demographic baseline, ensuring temporal alignment between the two sources. All analyses are conducted at the census tract level. The two outcomes of interest are the prevalence of self-reported poor mental health (≥14 days in the past month) and obesity prevalence among adults aged 18 and older.

For the ACS, I downloaded detailed table (B-table) variables rather than subject table (S-table) pre-computed estimates. This choice was deliberate: B-tables provide raw counts with explicit numerators and denominators, allowing me to compute proportions with full control over denominator definitions. Subject tables report pre-aggregated rates whose denominators are not always consistently defined across indicators. All proportions were computed from raw counts in the cleaning step. Variable selection rationale is described in Section 2.

After downloading the full PLACES dataset and filtering to the two outcome measures, the resulting file contained 78,815 tracts with no missing values on either outcome. For the ACS, data were downloaded for all 50 states and Washington D.C., yielding 73,056 tracts. Two variables exhibited Census-suppressed values: median household income (B19013, 1,024 NAs) and median gross rent as a percentage of income (B25071, 1,683 NAs). Inspection of the geographic distribution of missing values showed that both were dispersed across 48 states with no systematic geographic concentration, consistent with small-sample suppression rather than structural missingness. Critically, 974 of the 1,024 tracts missing median income were also missing rent burden (B25071) — a 95% overlap suggesting these are the same small-population tracts triggering suppression on multiple variables simultaneously. These tracts were excluded, leaving 71,323 ACS tracts.

A small number of additional NAs arose during proportion computation when denominators equaled zero — that is, tracts with no relevant population for a given indicator. Four variables were affected: public assistance rate (127 NAs), single-parent household rate (8 NAs), less-than-high-school rate (1 NA), and less-than-bachelor's rate (1 NA). Following the Gladish et al. (2026) reADI convention, these were coded as zero rather than excluded, as a rate of zero is substantively meaningful when no relevant population exists in a tract (e.g., no households with children under 18 for the public assistance rate denominator).

After merging the ACS and PLACES datasets on the 11-digit census tract FIPS code, 55,534 tracts were retained. The 15,789 unmatched tracts were distributed across 49 states with no evidence of systematic geographic bias, suggesting the mismatches likely reflect differences in

tract definitions across datasets rather than selective exclusion of particular community types, although this cannot be directly verified with the current data.

2. Index Construction: Reasoning & Process

Starting point: theory versus data

The core decision in constructing a neighborhood disadvantage index is whether to let theory or data drive variable selection. A purely data-driven approach — feeding all available ACS variables into a dimensionality reduction algorithm — risks including variables that are statistically related to health outcomes but conceptually unrelated to neighborhood disadvantage. Forthman et al. (2021), for example, applied factor analysis to a broad set of ACS variables and found that four of their five extracted factors primarily captured demographic composition — racial/ethnic makeup and age structure — rather than material deprivation, producing dimensions that predict mental health outcomes but do not constitute a coherent measure of neighborhood disadvantage. A purely theory-driven approach, on the other hand, risks overlooking dimensions that, while not prominent in existing frameworks, prove empirically important for predicting the outcomes of interest.

I adopted a hybrid strategy: use established theoretical frameworks to define the domain of relevant variables, then apply factor analysis to extract the latent structure within that domain.

Defining the theoretical dimensions

I began by reviewing four published neighborhood disadvantage indices—the Area Deprivation Index (ADI; Singh, 2003; Kind et al., 2014), the Social Deprivation Index (SDI; Butler et al., 2013), and the Reproducible Area Deprivation Index (reADI; Gladish et al., 2026)—alongside Sampson and Raudenbush's (1999) theoretical framework on concentrated disadvantage. I drew on Sampson & Raudenbush (1999) for its theoretical contribution to neighborhood disadvantage. Although these measures differ in name, scope, and data source, they share a common theoretical aim — capturing the multidimensional nature of neighborhood-level material and social deprivation.

Across these frameworks, six dimensions emerged consistently: economic deprivation, employment, education, family structure, housing conditions, and material resources. These dimensions reflect distinct but interrelated pathways through which neighborhood context shapes health. Table 1 summarizes the six dimensions and their constituent variables.

Table 1. Variables Across Five NDI-related Frameworks

Dimension	Variable	ADI - Singh (2003)	ADI - Kind (2014)	SDI - Butler (2013)	ReADI - Gladish et al. (2026)
Economic / Income	Poverty Rate	✓	✓	✓	✓
	150% Poverty threshold	✓	✓		
	Public assistance				✓
	Median household income	✓	✓		✓
	Income disparity	✓	✓		
Employment	Unemployment rate	✓	✓	✓ (considered)	✓
	Nonemployment rate		✓		
	White collar occupation (%)	✓	✓		
Education	< 9 years of schooling	✓	✓		
	No high school diploma (< 12y)	✓	✓	✓	✓
Family Structure	Single-parent households (%)	✓	✓	✓	✓
Housing	Renter-occupied housing (%)			✓	✓
	Housing crowding	✓	✓	✓	✓
	Median home value / rent / mortgage	✓	✓		
	Home ownership rate	✓	✓		
	Lacking complete plumbing	✓	✓		
Resources	No vehicle households (%)	✓	✓	✓	✓
	No telephone households (%)	✓	✓		

Note: *Empty cells indicate the dimension was not explicitly addressed in that framework. Sampson and Raudenbush (1999) is included primarily for its theoretical contribution on collective efficacy and social disorganization, rather than as a source of directly operationalizable variables.*

Variable selection within dimensions

To identify candidate ACS variables for each dimension, I systematically reviewed all 619 distinct B-table concepts in the ACS 2015–2019 5-year estimates, rather than importing variable lists directly from existing indices. This was necessary for two reasons. Several of the reference indices were built from non-ACS sources — the 1990 and 2000 Decennial Census (Singh, 2003; Kind et al., 2014) and the PHDCN survey (Sampson & Raudenbush, 1999) — meaning their specific variable operationalizations do not map cleanly onto ACS table structure. Even for indices originally using ACS data, my cross-index review identified theoretically supported dimensions that were absent or underrepresented in any single framework, requiring me to search beyond what existing indices had operationalized. I selected detailed tables (B-tables) over subject tables (S-tables) throughout, as described in Section 1.

For the economic dimension, I retained all four variables despite their high intercorrelations ($r = 0.74–0.85$ among the three poverty-related indicators) because each captures a distinct aspect of economic deprivation: poverty status reflects official threshold-based deprivation, SNAP receipt reflects food insecurity and government dependency, and public assistance captures cash transfer reliance. Median income, though negatively correlated with the others ($r = −0.64$ to $−0.67$), provides complementary information about the community's overall wealth level.

For the education dimension, I considered two thresholds: less than high school diploma and less than bachelor's degree. Rather than choosing one a priori, I computed both and retained both, motivated by the reADI's observation that in the contemporary U.S. labor market, a bachelor's degree increasingly represents the threshold for economic security (Gladish et al., 2026). Their moderate intercorrelation ($r = 0.64$) confirmed that the two variables capture different aspects of educational disadvantage.

For housing, I chose lacking kitchen facilities (B25052) over lacking plumbing (B25048) as the more meaningful indicator of contemporary material deprivation, given that plumbing deficits are extremely rare in modern U.S. housing stock. Median owner-occupied home value (ACS B25077) and owner-occupancy rate (derived from ACS B25003) were considered but dropped: home value exhibited comparable suppression rates to rent burden, and owner-occupancy rate is collinear with renter rate by construction, as the two sum to 100% of occupied units. Rent burden — defined as median gross rent as a percentage of household income (B25071) — was included in place of median owner-occupied home value (B25077) and owner-occupancy rate (B25003). Rent burden captures housing-induced financial stress more directly than asset-based measures: households that are cost-burdened have fewer residual resources for food, healthcare, and other necessities, linking housing disadvantage to the health outcomes of interest through a material deprivation pathway rather than a wealth accumulation pathway.

For the resources dimension, I retained no-vehicle rate (B08201) as a well-established indicator of geographic mobility constraints. I also included no-internet-access rate (B28003, defined as households lacking broadband subscription plus households with no computer), acknowledging this as a departure from classical NDI construction. Emerging evidence supports broadband access as a social determinant of health relevant to both outcomes of interest. Prior research suggests that areas with lower broadband access tend to have reduced availability of mental health providers (Kohli et al., 2024) and that lower broadband connectivity has been associated with higher prevalence of cardiometabolic conditions, including obesity (DeGrace et al., 2025). I

excluded telephone access following reADI guidance that telephone ownership no longer meaningfully discriminates neighborhood disadvantage in the contemporary U.S. context.

One variable was excluded post-hoc based on factor analysis results: lacking complete kitchen facilities loaded at only 0.158 on the primary factor, well below the conventional 0.3 threshold, suggesting it does not share sufficient common variance with the other indicators. A loading threshold of 0.30 was used as a minimum criterion for interpretability, consistent with common rules of thumb in exploratory factor analysis (e.g., Hair et al., 2010). Its removal improved the proportion of variance explained by the single factor from 49.0% to 52.5%.

Table 2. ACS Variables by Dimension

Dimension	ACS Table	Indicator	Rationale
Economic/Income	B17001	Poverty rate	Included in ADI, SDI and ReADI
	B19013	Median household income	Included in ADI and ReADI
	B22003	SNAP receipt rate	(*) Captures food insecurity more directly than public assistance rate; reflects inability to afford adequate nutrition
	B09010	Public assistance rate	Included in ReADI
Employment	B23025	Unemployment rate	Included in ADI, SDI and ReADI
Education	B15003	Less than HS diploma rate	Included in ADI, SDI and ReADI
	B15003	Less than bachelor's degree rate	Included in ReADI
Family Structure	B11003	Single-parent household rate	Included in ADI, SDI and ReADI
Housing	B25003	Renter-occupied rate	Included in SDI and ReADI
	B25014	Crowding rate	Included in ADI, SDI, and ReADI
	B25052	Lacking kitchen facilities rate	(*) Replaces lacking complete plumbing (B25048) as a marker of extreme housing inadequacy; retains

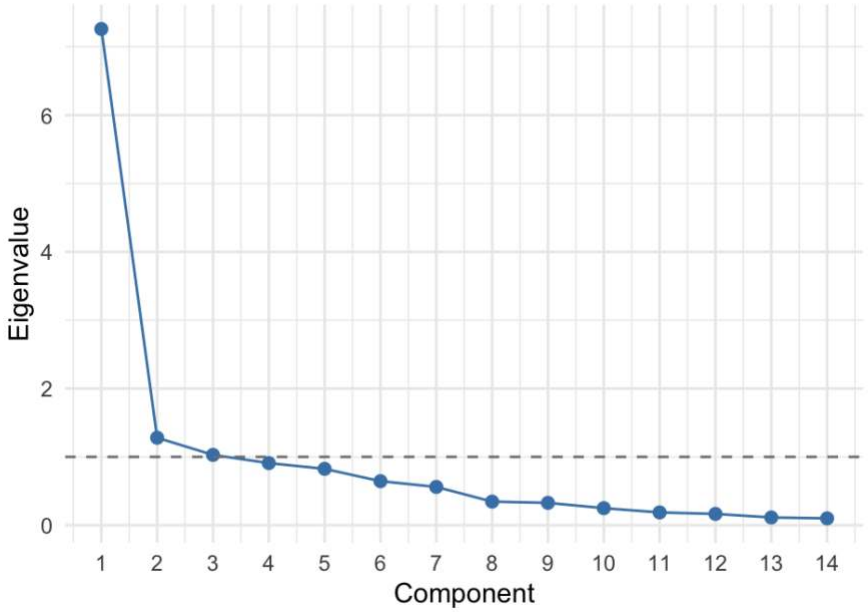
			relevance to food access and obesity risk; excluded after factor analysis
	B25071	Rent burden	(*) Replaces median owner-occupied home value (B25077) and owner-occupancy rate (derived from B25003) used in prior indices; shifts measurement from asset accumulation to housing-induced cash flow stress, capturing residual income constraints on food and healthcare
Resources	B08201	No-vehicle rate	Included in ADI, SDI and ReADI
	B28003	No internet access rate	(*) Novel indicator of digital deprivation; supported by empirical evidence linking lack of broadband access to reduced mental health service availability and higher cardiometabolic risk (Kohli et al., 2024; DeGrace et al., 2025)

Note: *Variables marked with (*) have no precedent in the five reference indices reviewed in Table 1 and were selected based on theoretical and empirical justification described in "Variable selection within Dimensions"*

Choosing the construction method

I considered three approaches to aggregating the selected variables: a simple additive scale, principal component analysis (PCA), and factor analysis (FA). I rejected the additive scale because it assigns equal weight to all variables regardless of their empirical relationship to the underlying construct. Between PCA and FA, I chose FA because my goal was to estimate a latent construct — neighborhood disadvantage — rather than to summarize variance in the observed data. FA explicitly models the shared variance among indicators as attributable to a common factor, which is more appropriate when indicators are theoretically understood as imperfect measurements of an underlying dimension.

Figure 1. Scree Plot of Principal Components



Before committing to a single-factor solution, I used PCA to assess whether the data supported this assumption. The first principal component explained 51.9% of variance with an eigenvalue of 7.26, with only three components exceeding eigenvalue 1.0. This pattern supported a dominant single-factor structure. I proceeded with single-factor maximum likelihood FA, which yielded loadings ranging from 0.365 to 0.923, with median household income loading negatively as expected (−0.896), confirming that the factor captures disadvantage rather than advantage. Figure 2a presents loadings ordered by magnitude to highlight which variables contribute most strongly to the factor, while Figure 2b groups the same loadings by theoretical dimension to illustrate how each dimension is represented in the index.

Figure 2a. Factor Loadings: Neighborhood Disadvantage Index

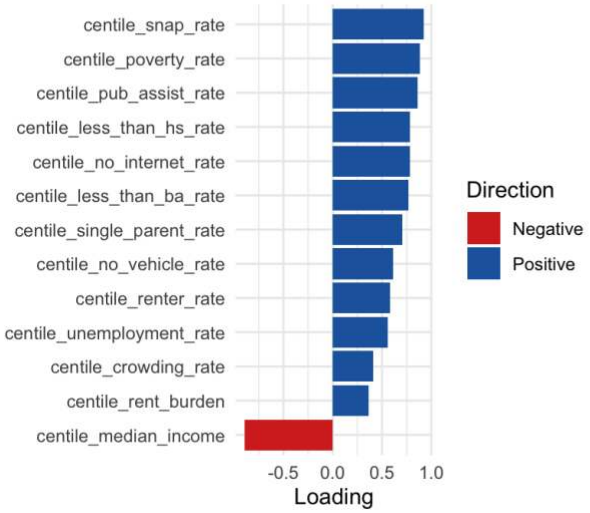
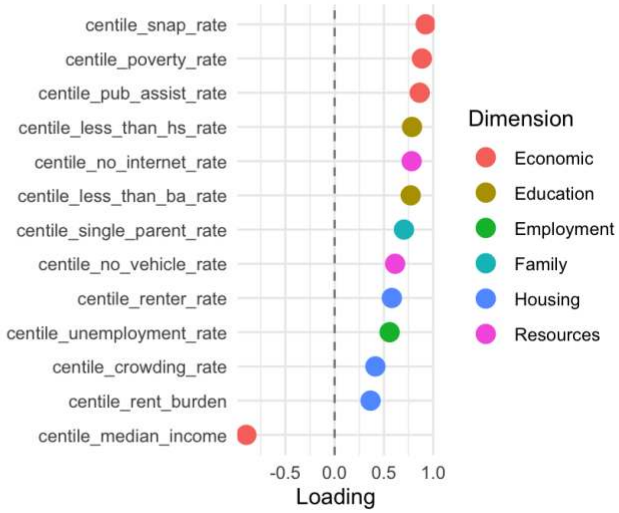


Figure 2b. Factor Loadings by Dimension



Following Gladish et al. (2026), factor scores were extracted directly from the FA output (using maximum likelihood estimation via psych::fa()) rather than applying manually derived weights to raw variables, avoiding the standardization errors identified in the NA-ADI. Gladish et al. also applied population weighting in their factor analysis to account for substantial variation in unit population size across block groups. I did not adopt this procedure at the census tract level, as tracts are explicitly designed to approximate population homogeneity (typically 2,500–8,000 residents), making between-unit size disparities substantially smaller than at the block group or county geographies that motivate the weighting approach.

Choosing the transformation method

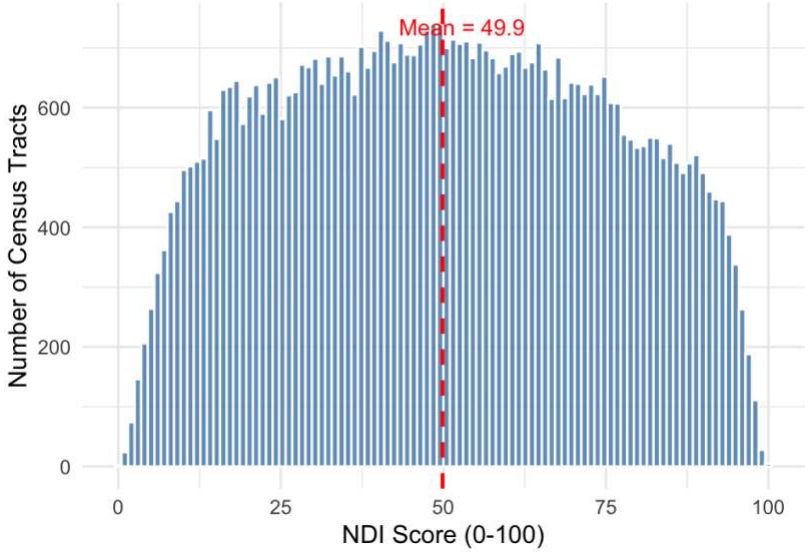
Prior to factor analysis, variables required transformation to address skewness and ensure comparability. I compared two approaches: log transformation followed by z-score standardization, and conversion to centile rankings. Following Butler et al. (2013), who found no substantive difference between these approaches in their SDI construction, I tested both. PCA on centile-ranked variables explained slightly more variance on the first component (51.9% vs. 49.6%) and yielded fewer components with eigenvalue above 1 (3 vs. 4), indicating a marginally cleaner single-factor structure. I therefore selected centile ranking as the primary transformation, retaining log + z-score for robustness checking.

The final index was scaled to 0–100 for interpretability, with higher scores indicating greater neighborhood disadvantage.

3. Results

The Neighborhood Disadvantage Index (NDI) was constructed from 13 ACS indicators spanning six dimensions of neighborhood disadvantage. The index follows an approximately normal distribution across 55,534 census tracts (mean = 49.9, SD = 24.9, range = 0–100), with higher values indicating greater disadvantage. The distribution is approximately symmetric, though this is not mechanically guaranteed by the centile transformation applied to input variables.

Figure 3. Distribution of Neighborhood Disadvantage Index (NDI)
N = 55,534 census tracts | Mean = 49.9, SD = 24.9



Association with obesity prevalence

OLS regression of tract-level obesity prevalence on the NDI yielded a statistically significant positive association ($\beta = 0.200$, SE = 0.001, $p < 2\times10^{-16}$). Each one-unit increase in NDI was associated with a 0.200 percentage point increase in obesity prevalence. The model explained 43.7% of the variance in tract-level obesity prevalence ($R^2 = 0.437$).

Association with poor mental health prevalence

The association between NDI and poor mental health prevalence was similarly significant and positive ($\beta = 0.0945$, SE = 0.0004, $p < 2\times10^{-16}$). Each one-unit increase in NDI was associated with a 0.0945 percentage point increase in the prevalence of self-reported poor mental health. The model explained a substantially larger share of variance in mental health prevalence ($R^2 = 0.558$) than in obesity prevalence, suggesting that neighborhood disadvantage is a stronger predictor of mental health outcomes than of obesity at the tract level.

Table 3. OLS Estimates of the Association Between Neighborhood Disadvantage and Health Outcomes

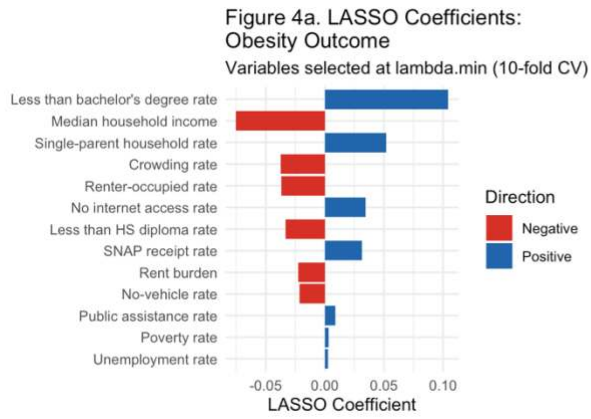
	(1) Obesity	**(2) Mental Health**
NDI (0–100)	0.200***	0.0945***
	(0.001)	(0.0004)
Constant	24.29***	12.34***
	(0.054)	(0.020)
Observations	55,534	55,534
R^2	0.437	0.558
RMSE	5.651	2.098

Notes: *Standard errors in parentheses. All models estimated using OLS at the census tract level. NDI = Neighborhood Disadvantage Index, scaled 0–100. *** $p < 0.001$.*

Robustness checks

Two robustness checks were conducted. First, I replicated the primary analysis using an alternative NDI constructed from log-transformed and z-score standardized indicators rather than centile rankings. Results were substantively identical: $R^2 = 0.425$ for obesity and $R^2 = 0.565$ for mental health, confirming that the findings are not sensitive to the choice of transformation method.

Second, I estimated LASSO regression models using all 13 constituent indicators as separate predictors, with 10-fold cross-validation. LASSO R^2 exceeded that of the composite index (obesity: 0.658; mental health: 0.629), indicating that some predictive information is lost in dimensionality reduction. However, the instability of LASSO coefficients under multicollinearity — several housing and economic indicators received unexpected negative coefficients — itself motivates the use of a composite index. When constituent indicators are highly intercorrelated, as observed in the economic dimension ($r = 0.74$–0.85), individual regression coefficients become difficult to interpret. Factor analysis addresses this by extracting the shared variance into a single latent construct, yielding a more stable and theoretically coherent measure.



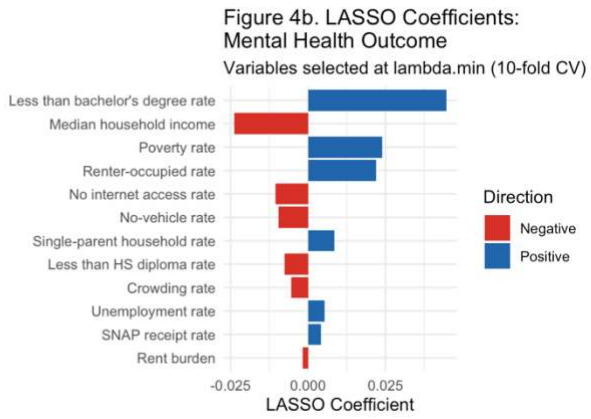


Table 4. Robustness Checks: Model Fit (R^2)

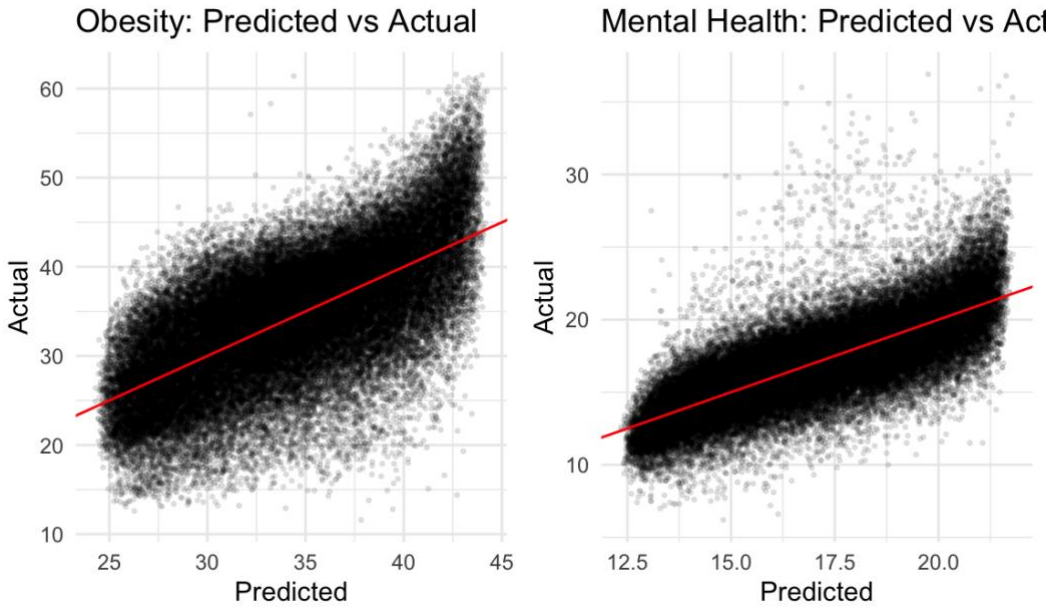
Model	Obesity R^2	Mental Health R^2
Primary (centile NDI)	0.437	0.558
Z-score NDI (RC1)	0.425	0.565
LASSO — 13 indicators (RC2)	0.658	0.629

Notes: *All models estimated at the census tract level. RC1 = Robustness Check 1 (z-score standardization); RC2 = Robustness Check 2 (LASSO with 10-fold cross-validation). R^2 for LASSO models calculated as 1 − (CV mean squared error / variance of outcome).*

Model validation

10-fold cross-validation of the primary models yielded R^2 and RMSE identical to in-sample estimates (obesity: CV R^2 = 0.437, RMSE = 5.65; mental health: CV R^2 = 0.558, RMSE = 2.10), confirming model stability and the absence of overfitting. The large sample size (n = 55,534) renders cross-validation results highly stable across folds. Predicted versus actual plots revealed no systematic bias across the range of predicted values for either outcome, though both models showed greater residual variance at lower predicted values, suggesting that neighborhood disadvantage is a less precise predictor for low-disadvantage tracts.

Figure 5. Predicted vs Actual Values: Primary OLS Models



4. Critical Reflection

The most consequential limitation of this analysis is the absence of spatial autocorrelation correction. Census tracts are geographically contiguous units, and health outcomes and socioeconomic conditions in neighboring tracts are likely to be correlated — a property known as spatial autocorrelation. Standard OLS regression assumes that residuals are independent across observations, an assumption almost certainly violated here. When this assumption is violated, standard errors are underestimated, p-values are artificially small, and the apparent precision of these estimates is overstated.

This limitation is not merely technical. The high R^2 values observed — particularly for mental health (0.558) — may in part reflect spatial clustering rather than a true causal relationship between neighborhood disadvantage and health outcomes. Tracts in the same metropolitan area share not only similar NDI scores but also similar healthcare infrastructure, environmental exposures, and historical disinvestment patterns that are not captured by the index. OLS regression cannot distinguish the contribution of neighborhood disadvantage per se from these spatially correlated confounders.

With more time and resources, I would address this limitation in two ways. First, I would apply spatial regression models — specifically a spatial lag model or spatial error model estimated via maximum likelihood — which explicitly account for the spatial dependence structure of the data. This would require constructing a spatial weights matrix defining which tracts are considered neighbors, most commonly based on shared boundaries (queen contiguity) or distance thresholds. Second, I would conduct spatial cross-validation, where training and test sets are geographically separated rather than randomly assigned, providing a more honest assessment of the model's ability to generalize across space rather than merely across adjacent tracts.

References

Butler, D. C., Petterson, S., Phillips, R. L., & Bazemore, A. W. (2013). Measures of social deprivation that predict health care access and need within a rational area of primary care service delivery. Health Services Research, 48(2 Pt 1), 539–559. https://doi.org/10.1111/j.1475-6773.2012.01449.x

DeGrace, S., Turk, N., Alhoch, Y., & Duru, O. K. (2025). Census-tract broadband connectivity and the risk of diabetes, hypertension and obesity/overweight in California. BMC Public Health, 25, 4341. https://doi.org/10.1186/s12889-025-25663-z

Forthman, K. L., Colaizzi, J. M., Yeh, H.-W., Kuplicki, R., & Paulus, M. P. (2021). Latent variables quantifying neighborhood characteristics and their associations with poor mental health. International Journal of Environmental Research and Public Health, 18(3), 1202. https://doi.org/10.3390/ijerph18031202

Gladish, N., Phillips, R. L., & Rehkopf, D. H. (2026). Estimation of mortality via the Neighborhood Atlas and Reproducible Area Deprivation Indices. JAMA Network Open, 9(1), e2546800. https://doi.org/10.1001/jamanetworkopen.2025.46800

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate data analysis (7th ed.). Pearson.

Kind, A. J. H., Jencks, S., Brock, J., Yu, M., Bartels, C., Ehlenbach, W., Greenberg, C., & Smith, M. (2014). Neighborhood socioeconomic disadvantage and 30-day rehospitalization: A retrospective cohort study. Annals of Internal Medicine, 161(11), 765–774. https://doi.org/10.7326/M13-2946

Kohli, K., Jain, B., Patel, T. A., Eken, H. N., Dee, E. C., & Torous, J. (2024). The digital divide in access to broadband internet and mental healthcare. Nature Mental Health, 2(1), 88–95. https://doi.org/10.1038/s44220-023-00176-z

Sampson, R. J., & Raudenbush, S. W. (1999). Systematic social observation of public spaces: A new look at disorder in urban neighborhoods. American Journal of Sociology, 105(3), 603–651. https://doi.org/10.1086/210356

Singh, G. K. (2003). Area deprivation and widening inequalities in US mortality, 1969–1998. American Journal of Public Health, 93(7), 1137–1143. https://doi.org/10.2105/AJPH.93.7.1137