# Enhancing Insights into Obesity and Health-Related Outcomes Using Regressions: A Study of the Carolinas

**Anthony Haitao Zou**[1] **& Alex Haopeng Liu**[2*]

***Received September 30, 2025***
***Accepted January 23, 2026***
***Electronic access February 15, 2026***

Poor physical and mental health, obesity, and physical inactivity are major public health concerns in the United States, with profound implications for population well-being and health system burden. In this study, we focus on four key county-level health outcomes, physically unhealthy days, mentally unhealthy days, percentage of adults with obesity, and percentage of physically inactive adults, to identify high-risk geographic regions and associated social determinants of health (SDOH). We integrate 300 county-level variables from the Agency for Healthcare Research and Quality's (AHRQ) SDOH Database, along with health outcome data from the County Health Rankings & Roadmaps. Multiple predictive linear regression modeling approaches, including ordinary least squares (OLS) regression, Lasso regression for modeling each response separately, and group Lasso regression for joint modeling of four responses, were applied to forecast the following year's values for each of the four outcomes. Model performance was evaluated using mean squared error, with Lasso and group Lasso regression achieving competitive results across outcomes compared to OLS regression. Key predictors include economic and social factors such as counties with a lower median home value of owner-occupied housing units and a higher percentage of family households with no spouse. The findings highlight specific geographic regions, such as inland rural areas, as areas of concern. These findings identify predictive associations that can guide hypothesis generation and preliminary policy discussions, while further causal analysis is needed before intervention.

**Keywords:** Behavioral and Social Sciences; County-Level Prediction; Mental and Physical Health Burden; Obesity; Physical Inactivity; Public Health

## Introduction

Chronic physical and mental health burdens, such as obesity, physical inactivity, and poor self-reported health, remain among the most pressing public health challenges in the United States[1–3]. Obesity alone is associated with increased risk for heart disease, diabetes, and certain cancers, contributing significantly to premature mortality and healthcare costs[4–6].

In parallel, physically and mentally unhealthy days, as well as insufficient physical activity, reflect broader well-being and functioning and are closely linked to underlying socioeconomic and environmental conditions[7,8]. These conditions are particularly concerning in the southeastern United States, where the prevalence of obesity and physical inactivity is often higher than the national average[9,10]. For instance, in 2024, either South Carolina (SC) or North Carolina (NC) report elevated rates of obesity (36% in SC and 36% in NC versus the national average of 34%) and poor mental health (average numbers of mentally unhealthy days reported in past 30 days (age-adjusted) are 4.5 in NC and 5.4 in SC versus the national average of 4.8) and physical unhealth days (average numbers of physically unhealthy days reported in past 30 days (age-adjusted) are 3.3 in NC and 3.8 in SC versus national average of 3.3) compared to national benchmarks[11]. These disparities underscore the urgent need for targeted, data-driven public health interventions.

It is documented that there are geographic variations in chronic disease burden (such as obesity[12], physical inactivity[13], self-reported poor physical and mental health[14]). Based on Behavior Risk Factors Surveillance System data (BRFSS) 1995-2012, Dwyer-Lindgren et al. reported substantial geographic disparities in poor self-reported health (SRH) including self-reported general health, physical distress, mental distress, and activity limitation[14]. Another study used both BRFSS and National Health and Nutrition Examination Survey (NHANES) demonstrated an increase in the prevalence of sufficient physical activity from 2001 to 2009 which was matched by an increase in obesity in almost all counties during the same time period[12].

During 1993–2001, BRFSS respondents in Alabama, Con-

[1] *River Bluff High School, 320 Corley Mill Road, Lexington, SC, 29072, USA*
[2] *Cary Academy, 1500 N Harrison Ave, Cary, NC 27513, USA*
**Corresponding author email: alexhliu30@gmail.com*

necticut, Maine, New Jersey, New Mexico, North Carolina, and Oregon reported both increasing physically and mentally unhealthy days [15]. An overall worsening physical and mental health were reported from January 1993 to December 2006 using monthly observed mean physically and mentally unhealthy days from BRFSS [16]. A significant trend over time for increasing fair/poor SRH was found based on NHANES 2001–2016 data [17].

Robust association was reported between the social determinants of health (SDOH) and SRH [7,18,19]. Scheinker, Valencia and Rodriguez found that county-level demographic, socioeconomic, health care, and environmental factors explain the majority of variation in county-level obesity prevalence [20]. Using 2000 BRFSS, Jia et al. indicated that socioeconomic variables predicted similar mean numbers of physical and mental unhealthy days at both the state and county level [21].

To better address these issues, it is essential to understand the spatial and temporal patterns of these outcomes and the SDOH that contribute to them. However, dynamic, county-level analyses of such health indicators remain limited, particularly those leveraging comprehensive SDOH data. This study aims to answer the following central question: "How can we dynamically predict county-level obesity, physical and mental health burden, and physical inactivity using SDOH data, and identify the most influential predictors?"

We use the Agency for Healthcare Research and Quality (AHRQ) SDOH Database, a publicly available, longitudinal resource that offers standardized county-level data across multiple domains of SDOH, including economic stability, education, healthcare access and quality, neighborhood environment, and community context [22]. We integrate this dataset with health outcome measures from the County Health Rankings & Roadmaps to model and forecast county-level trends [23].

The key contributions of this study include:

- Dynamic prediction of health outcomes using several linear regression methods, including ordinary least squares (OLS) regression, Lasso regression, and group Lasso regression;
- Incorporation of comprehensive SDOH features (300 variables); and
- Identification and interpretation of the most influential SDOH markers associated with each outcome.

Our work is novel in its integration of penalized regression and joint modeling of multiple outcomes combined with time-aware validation to enhance forecasting robustness. By identifying high-risk counties and critical social determinants, our findings can inform targeted public health strategies aimed at improving population health and reducing disparities.

## Methods

To forecast next-year outcomes for county-level obesity, physical inactivity, and physically and mentally unhealthy days, we employed a suite of linear regression models: OLS regression, Lasso regression, and group Lasso regression using comprehensive SDOH predictors. Based on the County Health Rankings & Roadmaps database, the obesity is measured by percentage of adults that report BMI $\geq$ 30; physical inactivity is reported as percentage of adults that report no leisure-time physical activity; the physically unhealthy days is defined as average number of reported physically unhealthy days per month; and mental health unhealthy days is calculated by the average number of reported mentally unhealthy days per month.

A comprehensive set of SDOH variables for counties in NC and SC were combined using the AHRQ SDOH Database. This curated resource integrates data from multiple federal and public datasets, including the American Community Survey (ACS), the Area Health Resources Files (AHRF), and the American Foundation for AIDS Research (AMFAR).

### Data Processing

To ensure consistency and completeness across both states, we restricted our analysis to variables with no missing data across all counties in NC and SC. This filtering process resulted in a final set of 300 predictor variables. The majority of the variables (260) were sourced from ACS, capturing detailed demographic and socioeconomic information. An additional 27 variables representing healthcare provider characteristics were obtained from AHRF, while 13 variables describing healthcare facility attributes were drawn from AMFAR.

For modeling purposes, the positive physically and mentally unhealthy days variables were log-transformed with the following formula:

$$x \to \ln x. \tag{1}$$

The proportion-based outcomes were logit transformed using the following formula:

$$x \to \ln \frac{x}{100 - x}. \tag{2}$$

These transformations convert the count or percentage outcome variables to a continuous scale on the real line without restriction, enabling more effective linear regression. Note that our dataset contains no zero counts and no percentages equal to 0 or 1, so the log transformation does not produce undefined values.

For the predictors, all 300 SDOH variables were standardized before analysis to place them on a common scale. Each variable was transformed to have a mean of zero and a standard deviation of one, ensuring comparability across features

with different units and magnitudes. This preprocessing step prevents any single variable from disproportionately influencing the model due to its scale, which is particularly important when using regularization-based methods[24].

## Procedures

To evaluate model performance, we used a year-based cross-validation framework in which each year from 2017 to 2021 was used for prediction evaluation while training was performed on all preceding years. The prediction error for each year was calculated using mean squared error (MSE), and we reported the average MSE across all folds for comparison.

To support the goal of forecasting future outcomes, we implemented a one-year temporal lag between predictors and response variables. For example, data from 2016 were used to predict health outcomes in 2017, data from 2017 to predict 2018 outcomes, and so on. This design mimics a real-world prediction scenario where only past and present information is available for future projections.

We adopted a 3:1:1 temporal split of the data into training, validation, and test sets, respectively. That is, we trained the models using predictors from 2016–2018 with outcomes from 2017–2019, then validated on the next year (2019 predictors, 2020 outcomes) and tested on the following year (2020 predictors, 2021 outcomes). This allocation ensures sufficient data for model learning while preserving dedicated subsets for hyperparameter tuning and final model evaluation. Table 1 summarizes the datasets used across the modeling pipeline. The training set comprises three years of data from 146 counties (438 by 304 observations with 300 predictors and 4 outcomes). Each validation and prediction set uses one year of data, with 146 by 304 observations.

A major challenge in this framework is the limited sample size, especially for the training and validation sets. With only a few years of county-level data, statistical power to detect meaningful associations is reduced. This limitation also makes model calibration more difficult and increases the risk of overfitting, particularly in the context of the high-dimensional predictor space (300 SDOH features). Careful model selection and validation were therefore critical to mitigate instability in performance estimation.

To prevent overfitting and make efficient use of the data, we employ a two-stage approach for methods requiring parameter tuning, including Lasso and group Lasso. Models are first trained on the training set, and tuning parameters are selected based on performance of the validation set. The final model is then refit using the combined training and validation data. For OLS regression, which does not require tuning, we fit the model directly on the combined training and validation set. Predictive performance is assessed on a separate test set.

### Linear Regression

Linear regression models the relationship between a continuous outcome variable and one or more independent variables, under the assumption of a linear relationship[25]. The model is expressed as:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \tag{3}$$

where $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the regression coefficients, $x_{i1}, x_{i2}, \ldots, x_{ip}$ are the predictor variables and $\varepsilon_i$ is the error term in the $i$-th county for $i = 1, \ldots, n$. The model is trained by minimizing the least squares cost function:

$$J(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2, \tag{4}$$

which measures the sum of squared differences between observed and predicted values.

### Lasso Regression

Lasso regression builds upon linear regression by adding an L1 regularization term to the cost function[26,27]. This penalizes the absolute values of the coefficients, encouraging sparsity in the model and enabling automatic variable selection. The Lasso cost function is:

$$J(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{5}$$

where $\lambda$ is the regularization parameter that controls the strength of the penalty. By shrinking less relevant coefficients toward zero, Lasso improves model interpretability and helps reduce overfitting. To determine the optimal level of regularization, we implemented a hyperparameter tuning procedure that evaluated a range of $\lambda$ values. For each value, we trained a Lasso model on the training set and evaluated its performance using MSE on a validation set. The model with the lowest validation MSE was selected as the best-performing Lasso model. In addition to overall prediction accuracy, we examined variable importance using the size of coefficients. Lasso selects a sparse set of predictors through penalized regression, enabling identification of key variables with strong linear associations.

### Group Lasso Regression

The OLS regression and Lasso regression are applied separately to each response variable. While this approach is useful, it presents challenges for interpreting results across multiple responses. To address this and gain insights into predictors that influence all outcomes, we aim to jointly model the response variables and identify variables that are important

**Table 1** Data used in each stage of modelling

| Stage | Sample Size | Predictors | Response |
|---|---|---|---|
| Training | 438 | SDOH Data from 2016–2018 | Four health outcomes in 2017–2019 |
| Validation | 146 | SDOH Data from 2019 | Four health outcomes in 2020 |
| Predictions | 146 | SDOH Data from 2020 | Four health outcomes in 2021 |

across all four[28,29]. Specifically, we estimate four regression models, one for each response, and treat the corresponding coefficients for each predictor as a group. The group Lasso penalty promotes group-wise sparsity: if a variable is unimportant for all responses, the entire group of coefficients is shrunk to zero. This encourages the selection of predictors that are jointly relevant, resulting in a more interpretable and effective model for our multivariate prediction task.

To further illustrate the idea, we use $k$ to denote the $k$-th response variable. Then the model can be written as

$$y_{ik} = \beta_{0k} + \sum_{j=1}^{p} \beta_{jk} x_{ij} + \varepsilon_{ik}, \tag{6}$$

where $\beta_{0k}, \beta_{1k}, \ldots, \beta_{pk}$ are the regression coefficients for the $k$-th response, $x_1, \ldots, x_p$ are the predictor variables as before, and $\varepsilon_{ik}$ is the error term for the $k$-th model for $k = 1, \ldots, K$. In our case, $K = 4$. Then, we apply the group Lasso regression for our problem. We define $p$ different groups for $p$ variables, each group of size $K$ corresponding to $K$ response variables. The group Lasso imposes an L2-norm for each group to impose groupwise sparsity. Thus, the corresponding objective function to minimize is:

$$J(\beta) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n} \left( y_{ik} - \beta_{0k} - \sum_{j=1}^{p} \beta_{jk} x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \sqrt{\sum_{k=1}^{K} \beta_{jk}^2}. \tag{7}$$

The joint linear model layout can be illustrated as follows:

$$\begin{bmatrix} X & 0 & 0 & 0 \\ 0 & X & 0 & 0 \\ 0 & 0 & X & 0 \\ 0 & 0 & 0 & X \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}, \tag{8}$$

where $X$ is the $n$ by $p$ data matrix. Then the groups can be shown as

$$[\beta_1 \quad \beta_2 \quad \beta_3 \quad \beta_4] = \begin{bmatrix} \beta_{11} & \beta_{12} & \beta_{13} & \beta_{14} \\ \vdots & \vdots & \vdots & \vdots \\ \boxed{\beta_{j1}} & \boxed{\beta_{j2}} & \boxed{\beta_{j3}} & \boxed{\beta_{j4}} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{p1} & \beta_{p2} & \beta_{p3} & \beta_{p4} \end{bmatrix}. \tag{9}$$

Once the group Lasso solution is obtained, we identify the nonzero groups for the final selected set of predictors.

### Trend Analysis and Visualization

We used pie charts, pairwise matrix plots, scatterplots, bar charts, and heatmaps to examine distributions, correlations, and spatiotemporal patterns across the four outcomes and to summarize prediction results. To assess relationships among outcomes, we built a pairwise matrix: diagonal panels show each outcome's marginal distribution (bar charts), and off-diagonal panels show pairwise scatterplots with overlaid Pearson correlations to quantify association strength and direction. To characterize spatial patterns, we generated geographic heatmaps for each year, mapping county values to a color scale (darker = higher). Sequencing yearly heatmaps enabled visual comparison of regional variability and temporal change. After model selection, we applied the same heatmap approach to highlight important predictors across counties and to compare observed versus model-predicted outcomes.

## Results

Figure 1 presents the distribution of 300 SDOH variables across 14 topics. Demographics accounts for the largest share, followed by housing and health insurance, underscoring the central role of structural and social factors in shaping population health outcomes. A breakdown of the 300 variables by data source is provided in the Supplementary file. Demographics comprised 94 variables, capturing population structure such as age and household composition. Housing (34 variables) and health insurance status (33 variables) reflected affordability, housing characteristics, and insurance coverage types. Employment (29 variables) and poverty (23 variables) described labor force participation and income relative to federal poverty thresholds. Measures of healthcare access included healthcare providers (14 variables) and healthcare facilities (13 variables), while income (14 variables) captured economic resources and inequality. Transportation, living conditions, and educational attainment each included 10 variables, reflecting commuting patterns, family structure, and educational levels. Smaller categories included immigration (9 variables), disability (4 variables), and migration (3 variables),

capturing citizenship status, disability prevalence, and residential mobility.

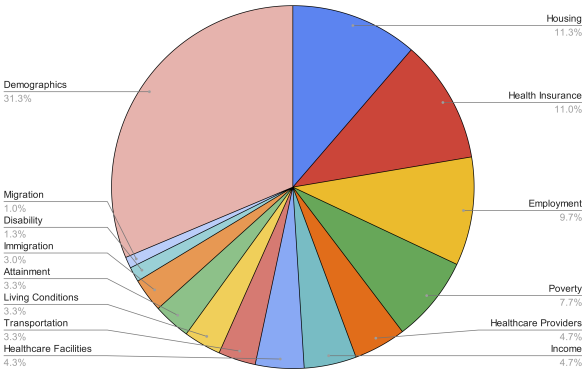


**Fig. 1** Pie chart of the 300 selected SDoH predictor variables by 14 topics.

We analyzed four county-level health outcomes from the County Health Rankings and Roadmaps dataset: the average number of physically unhealthy days reported per month, the average of mentally unhealthy days reported per month, the percentage of adults classified as obese, and the percentage reporting physical inactivity.

To explore the relationships among these outcomes, we present a matrix of pairwise scatterplots for the 2017 data in Figure 2. The patterns for other years are similar and not included here. The plots reveal relatively strong linear associations, with correlation coefficients of at least 0.6. This aligns well with intuition: higher numbers of physically and mentally unhealthy days, along with higher rates of physical inactivity, are naturally associated with higher obesity prevalence. These findings underscore the value and relevance of jointly modeling the four outcomes.

Physically unhealthy days: the average number of physically unhealthy days reported per month; Mentally unhealthy days: the average of mentally unhealthy days reported per month, % obese: the percentage of adults classified as obese in percent, and % Physically inactive: the percentage reporting physical inactivity in percent.

To examine outcome trends, we present the average outcome values over time (2017-2021) for the top 10 counties in terms of each outcome in 2017 in Figure 3. Note, all outcomes are expressed as population-normalized measures (averages or percentages), thereby minimizing sensitivity to county population size. While some fluctuations are observed, the overall patterns show a steady increase over the years for outcomes of Mentally Unhealthy Days and % of Obesity. For example, % of obesity in Fairfield and Chester counties has a clear increasing trend during 2018-2021.

Figures 4-7 present county-level heatmaps of four outcomes. The heatmaps reveal consistent spatial patterns across years and an overall worsening trend over time. To further characterize temporal patterns, we quantified the consistency of changes over time for each outcome by counting (i) counties exhibiting a strictly monotonic increase across all four years and (ii) counties showing increases in three of the four years. For obesity prevalence, 35 counties demonstrated monotonic increases, and 56 increased in three of four years. Physical inactivity showed fewer strictly monotonic increases (16 counties), though 70 counties increased in three of four years. No counties exhibited strictly monotonic increases in either physically unhealthy days or mentally unhealthy days; however, 13 and 74 counties, respectively, showed increases in three of the four years. York County has demonstrated the consistent increase in both obese and physical inactivity. Edgefield, Lee, and Macon Counties show a consistent increasing pattern in three of the four years across all four outcomes. 19 counties have consistent increasing pattern in three of the four years among three outcomes.

Table 2 reports the prediction accuracy for the four outcomes in 2021. Across all outcomes, transformed predictors consistently yield substantially lower MSEs than untransformed predictors, indicating improved model performance after transformation. Among the three modeling approaches, except mental unhealthy days, Lasso generally produces the lowest MSE for transformed predictors, followed by Group Lasso, and OLS gives the worst performance though differences between OLS and Group Lasso are relatively small for obesity. Interestingly, OLS tends to perform slightly worse than the penalized methods after transformation, although it works the best for the outcome of mental unhealthy days. Although Lasso often achieved the lowest MSE among transformed predictors, group Lasso was chosen for subsequent analyses because it allows for structured feature selection by groups of related predictors, improving interpretability and aligning with the substantive domains of social determinants of health. This approach facilitates more meaningful recommendations for policy and intervention, as results can be interpreted at the domain level rather than focusing solely on individual covariates.

We further demonstrate the most important features based on the group lasso approach. In particular, we calculate the group norm for each variable, i.e., $\sqrt{\sum_{k=1}^{K} \beta_{jk}^2}$ for variable $j$ and then sort them. Figure 8 presents the top 10 most important variables using the group norm, including factors such as the median home value of owner-occupied housing units. These factors are most associated with the four outcomes. To further understand the direction of associations for each factor with the outcomes, Figure 9 shows heatmaps of the most important predictors and corresponding outcomes for the top 5 (right 5 columns) and bottom 5 counties (left 5 columns)

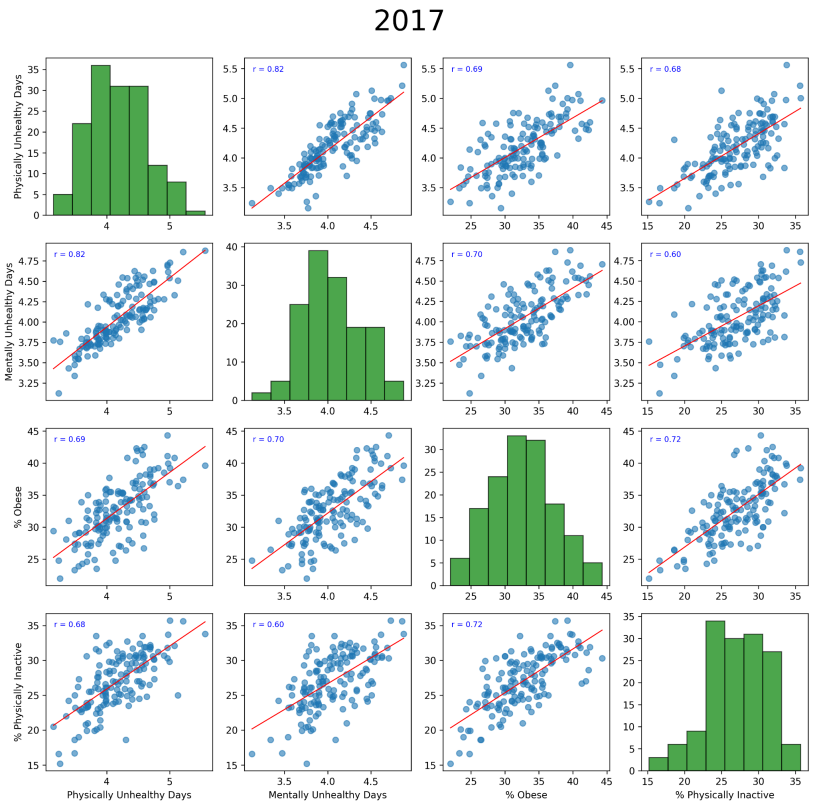


**Fig. 2** Matrix scatterplots among four outcome variables for the year 2017. The diagonal panels show histograms of the four outcomes (x-axis: variable range; y-axis: frequency). The off-diagonal panels show pairwise scatterplots of the outcomes, with the Pearson correlation and the best-fit linear regression line overlaid (x-axis and y-axis represents two predictors with their range).

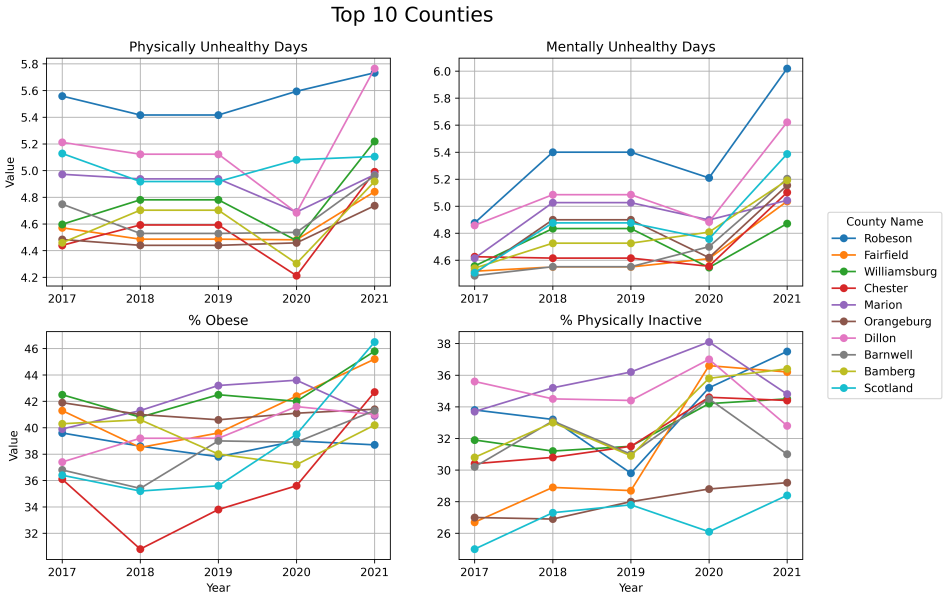


**Fig. 3** Trends for the average four outcomes for top 10 counties in terms of 2017 ranking during 2017-2021.

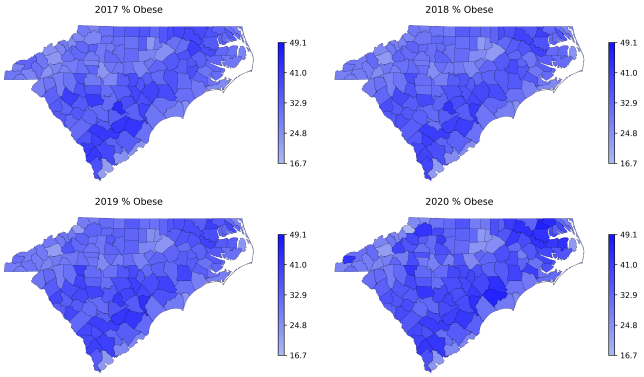


**Fig. 4** Color maps for % of obesity in Carolinas during 2017-2020

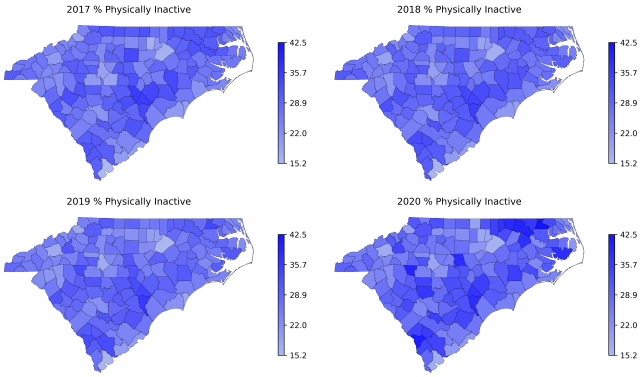


**Fig. 5** Color maps for % of physically inactive days in Carolinas during 2017-2020

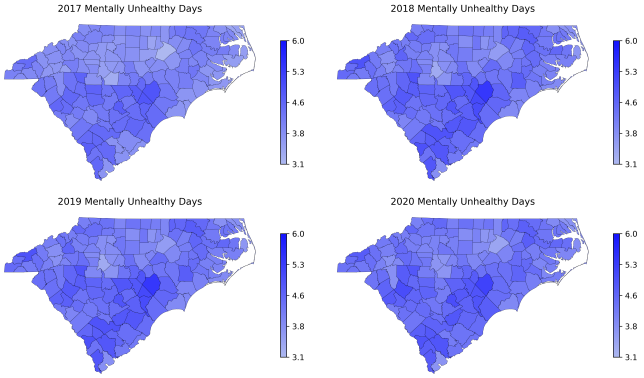


**Fig. 6** Color maps for mentally unhealthy days in Carolinas during 2017-2020

sorted by the average outcomes. The results indicate that counties with higher median home values tend to have lower obesity rates, lower percentages of physical inactivity, and fewer mentally and physically unhealthy days. In contrast,

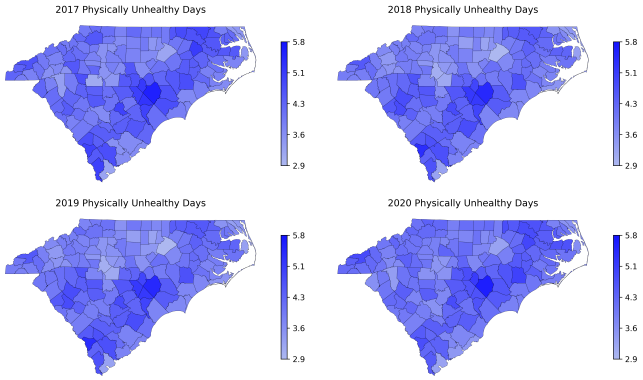


**Fig. 7** Color maps for physically unhealthy days in Carolinas during 2017-2020

**Table 2** Prediction accuracy for four outcomes based on three regression models: OLS, Lasso and group Lasso

| Outcome | | OLS | Lasso | Group Lasso |
|---|---|---|---|---|
| Obesity | Transformed | 0.04507 | 0.02914 | 0.04417 |
| | Untransformed | 23.12098 | 14.72717 | 22.73726 |
| Physically unhealthy | Transformed | 0.01762 | 0.01047 | 0.01680 |
| | Untransformed | 0.35889 | 0.20727 | 0.29761 |
| Mentally unhealthy | Transformed | 0.01690 | 0.02842 | 0.02940 |
| | Untransformed | 0.34858 | 0.57479 | 0.58746 |
| Physically Inactive | Transformed | 0.04683 | 0.02865 | 0.03436 |
| | Untransformed | 17.83918 | 11.31083 | 13.89089 |

higher percentages of family households without a spouse are associated with higher obesity rates and greater numbers of unhealthy days. Finally, Figure 10 maps the true and predicted outcomes for 2021, showing that the predictions capture spatial patterns well. However, the predicted heatmaps are generally lighter than the observed ones, suggesting underestimation. Two factors may contribute: (1) ongoing worsening of these outcomes over time that the models do not fully capture, and (2) limited sample size, which constrains predictive accuracy.

## Discussion

This study demonstrates the value of integrating comprehensive SDOH data with advanced regression-based modeling to forecast county-level health outcomes: obesity, physical and mental health burden, and physical inactivity. Using the AHRQ SDOH Database with outcome measures from County Health Rankings & Roadmaps, we compared OLS, Lasso, and Group Lasso on both raw and transformed predictors. Prepro-

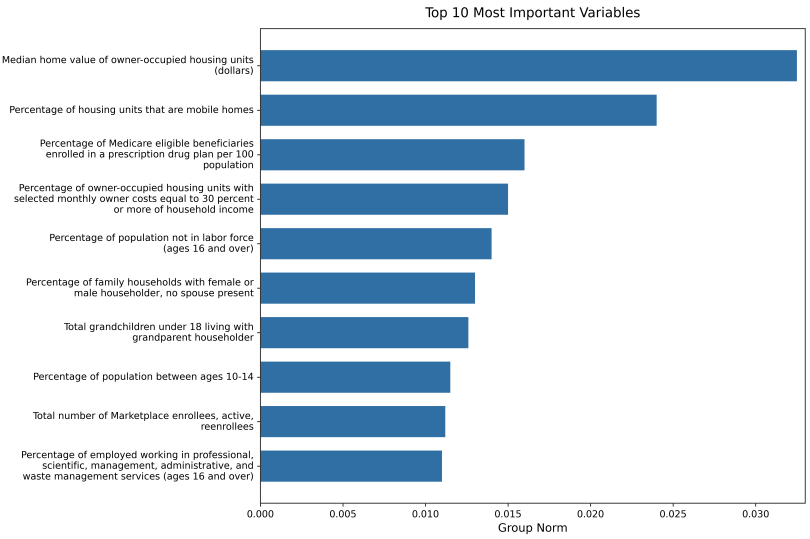


**Fig. 8** Top predictors sorted by the group norm of coefficients by group Lasso

cessing transformations substantially improved accuracy, especially for outcomes with wide variation in scale. Penalized methods generally work well.

Among the methods, Group Lasso was selected for subsequent analyses because it performs structured feature selection at the domain level. This is well suited to county-level inference, where SDOH variables cluster naturally into domains (e.g., socioeconomic status, healthcare access, built environment). Group-level selection enhances interpretability by highlighting the collective contribution of domains rather than isolated variables, which in turn supports clearer communication and more actionable, domain-targeted policy recommendations[30].

Beyond model performance, incorporating 300 SDOH variables provided a richer view of socioeconomic, environmental, and community factors associated with adverse outcomes. The combination of high-dimensional, longitudinal data and penalized regression offers a scalable framework for flagging high-risk counties and supporting dynamic public-health surveillance to inform equity-focused planning.

The selected predictors should be interpreted as predictive correlates, not necessarily causal drivers[31]; policy implications should not be drawn from these associations without additional causal analysis that addresses confounding, measurement error, and other potential biases.

## Conclusion

Our study combines the AHRQ SDOH Database with outcome measures from County Health Rankings & Roadmaps and applies OLS, Lasso, and Group Lasso regression. We find that transforming predictors and using penalized methods, especially Group Lasso, provides reasonable out-of-sample prediction of county-level health outcomes and highlights domain-level associations that can guide further investigation and, with appropriate caution, inform policy and intervention planning.

## Acknowledgments

The authors would like to thank Dr. Peiyin Hung at the University of South Carolina for mentoring, the helpful comments and suggestions from the review team, the AHRQ for developing the AHRQ SDoH Database, and the University of Wisconsin Population Health Institute for County Health Ranking & Roadmaps data, and for making them available for public use.

## References

1 D. R. Brown, D. D. Carroll, L. M. Workman, S. A. Carlson and D. W. Brown, *Quality of Life Research*, 2014, **23**, 2673–2680.
2 M. Bayliss, R. Rendas-Baum, M. K. White, M. Maruish, J. Bjorner and S. L. Tunis, *Health and Quality of Life Outcomes*, 2012, **10**, 154.
3 E. Robinson, A. Haynes, A. Sutin and M. Daly, *Obesity Science & Practice*, 2020, **6**, 552–561.
4 G. A. Bray, *Endocrinology and Metabolism Clinics*, 2003, **32**, 787–804.
5 S. Sarma, S. Sockalingam and S. Dash, *Diabetes, Obesity and Metabolism*, 2021, **23**, 3–16.
6 D. Mohajan and H. K. Mohajan, *Journal of Innovations in Medical Research*, 2023, **2**, 12–23.
7 H. Jia, D. G. Moriarty and N. Kanarek, *Journal of Community Health*, 2009, **34**, 430–439.
8 K. C. McNamara, E. T. Rudy, J. Rogers, Z. N. Goldberg, H. S. Friedman, P. Navaratnam and D. B. Nash, *Population Health Management*, 2024, **27**, 307–311.

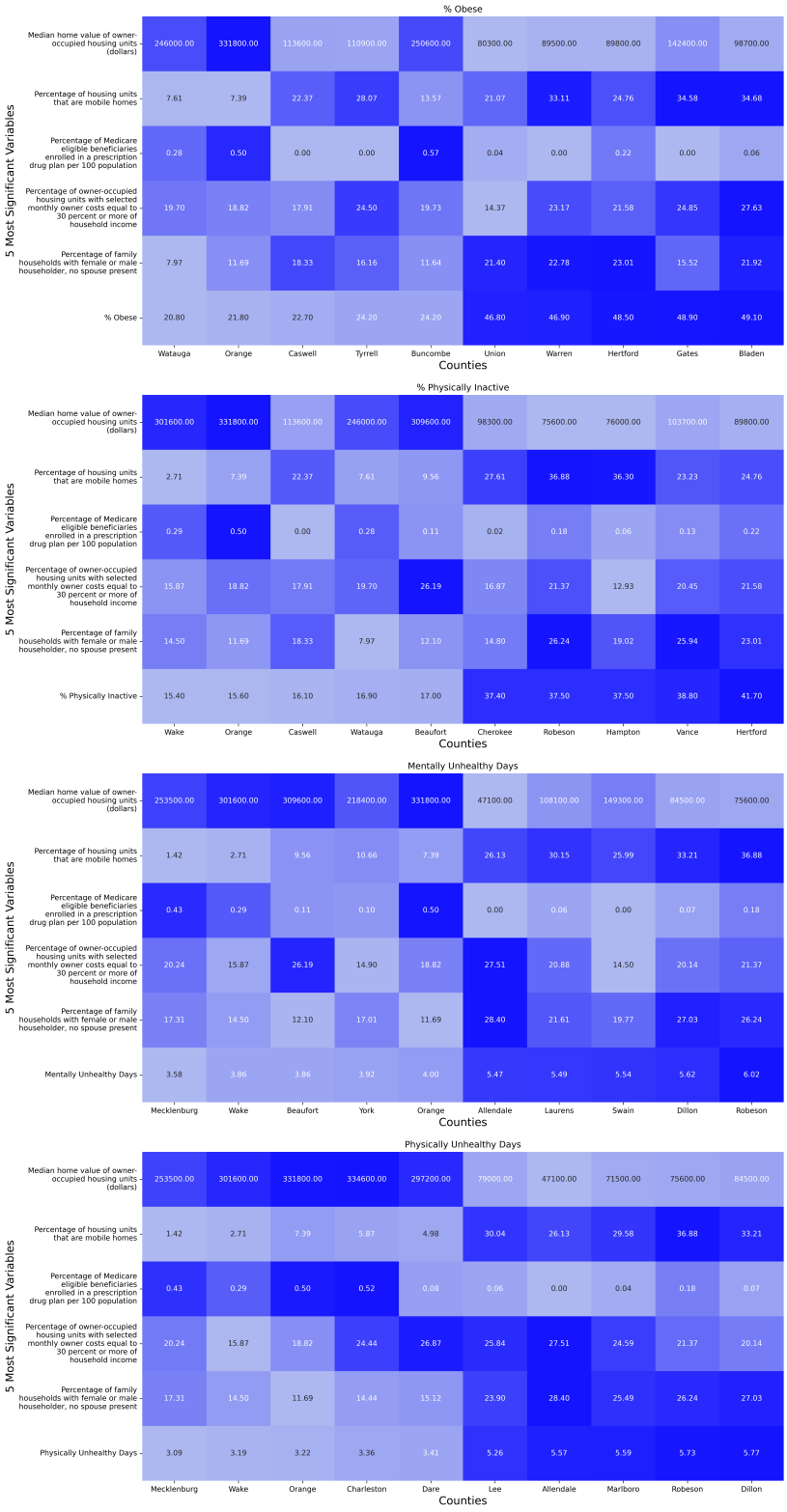


**Fig. 9** Heatmap of top 5 and bottom 5 counties by outcome variables and the 5 most important factors. Darker (lighter) colors indicate higher (lower) values.

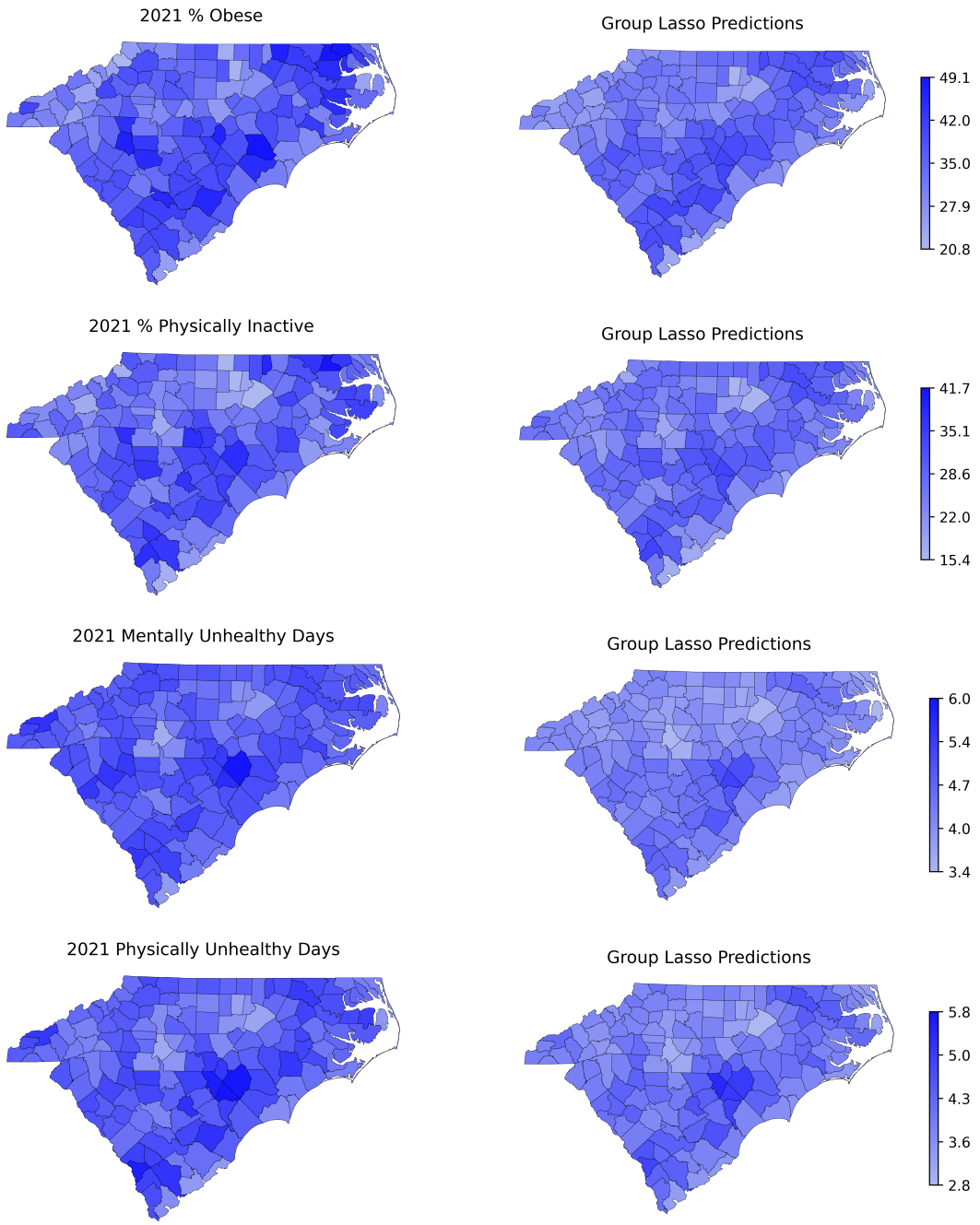


**Fig. 10** True and predicted values for four outcomes in 2021.


9 D. F. Terrell, *North Carolina Medical Journal*, 2002, **63**, 281–286.

10 D. G. Moriarty, M. M. Zack and R. Kobau, *Health and Quality of Life Outcomes*, 2003, **1**, 37.

11 University of Wisconsin Population Health Institute, *Compare Counties*, 2024, https://www.countyhealthrankings.org/health-data/compare-counties?year=2024&compareCounties=37000%2C45000%2C00000, Retrieved Dec 25.

12 N. C. Black, *Applied Spatial Analysis and Policy*, 2014, **7**, 283–299.

13 R. An, X. Li and N. Jiang, *International Journal of Environmental Research and Public Health*, 2017, **14**, 1326.

14 L. Dwyer-Lindgren, J. P. Mackenbach, F. J. van Lenthe and A. H. Mokdad, *Population Health Metrics*, 2017, **15**, 16.

15 H. S. Zahran, R. Kobau, D. G. Moriarty, M. M. Zack, J. Holt, R. Done-

hoo and Centers for Disease Control and Prevention, *MMWR Surveillance Summaries*, 2005, **54**, 1–35.
16 H. Jia and E. I. Lubetkin, *Public Health Reports*, 2009, **124**, 692–701.
17 M. L. Greaney, S. A. Cohen, B. J. Blissmer, J. E. Earp and F. Xu, *Quality of Life Research*, 2019, **28**, 3249–3257.
18 K. Wind, B. Poland, F. HakemZadeh, S. Jackson, G. Tomlinson and A. Jadad, *Health Promotion International*, 2023, **38**, daad165.
19 C. Delpierre, V. Lauwers-Cances, G. D. Datta, T. Lang and L. Berkman, *Journal of Epidemiology & Community Health*, 2009, **63**, 426–432.
20 D. Scheinker, A. Valencia and F. Rodriguez, *JAMA Network Open*, 2019, **2**, e192884.
21 H. Jia, P. Muennig, E. Lubetkin and M. R. Gold, *Journal of Epidemiology & Community Health*, 2004, **58**, 150–155.
22 U.S. Department of Health and Human Services, *Social Determinants of Health Database*, `https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html`, Retrieved April 25.
23 University of Wisconsin Population Health Institute, *County Health Rankings & Roadmaps, Health Data*, `https://www.countyhealthrankings.org/health-data`, Retrieved April 25.
24 J. M. H. Pinheiro, S. V. B. de Oliveira, T. H. S. Silva, P. A. R. Saraiva, E. F. de Souza, R. V. Godoy, L. A. Ambrosio and M. Becker, *arXiv preprint*, 2025.
25 S. Chatterjee and A. S. Hadi, *Regression analysis by example*, John Wiley & Sons, 2015.
26 J. Ranstam and J. A. Cook, *Journal of British Surgery*, 2018, **105**, 1348.
27 R. Tibshirani, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 1996, **58**, 267–288.
28 M. F. Duarte, W. U. Bajwa and R. Calderbank, *The performance of group Lasso for linear regression of grouped variables*, Duke University, Dept. Computer Science Technical Report TR-2010-10, 2011.
29 M. Yuan and Y. Lin, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2006, **68**, 49–67.
30 H. Ohanyan, L. Portengen, A. Huss, E. Traini, J. W. Beulens, G. Hoek, J. Lakerveld and R. Vermeulen, *Environment International*, 2022, **158**, 107015.
31 N. Altman and M. Krzywinski, *Nature Methods*, 2015, **12**, 899.