Regularized Regression Approaches to Understanding North Carolina Cancer Inequities

Madison E. Thompson

***Abstract*— *Objective:* This study evaluates how county-level social determinants of health (SDOH) are associated with breast, lung, and colorectal cancer incidence and mortality in North Carolina. *Methods:* Pooled ordinary least squares (OLS) and Ridge regression models were used to assess predictive capacity across all cancer outcomes. To identify outcome-specific predictors, six cancer–outcome models were fit using Least Absolute Shrinkage and Selection Operator (LASSO) regression. Model diagnostics, including variance inflation factors (VIF), residual analysis, and Cook's distance, were conducted to evaluate multicollinearity, model fit, and influential observations. *Results*: Pooled OLS models exhibited minimal explanatory power and violated key regression assumptions, supporting the use of regularized models. LASSO results indicated meaningful heterogeneity across cancer types, with higher education consistently appearing as a protective variable, and poverty and food insecurity more strongly associated with mortality measures. Several SDOH predictors demonstrated weak or inconsistent effects across cancers. *Conclusion*: These findings highlight the limitations of pooled linear modeling for cancer outcomes and demonstrate that specific SDOH factors differentially influence cancer incidence and mortality. This data may be relevant to public health policy in the State of North Carolina.**

***Index Terms*— cancer outcomes, LASSO, OLS, public policy, Ridge regression, social determinants of health**

I. INTRODUCTION

CANCER is the second-leading cause of death in the United States, and the leading cause of death in North Carolina [1], [2]. Emerging evidence suggests that epigenetics may mediate relationships between environmental exposures and cancer risk, presenting an interesting direction for understanding state- and county-level disease metrics [3]. While this paper cannot feasibly explore a full policy to molecular biology pipeline, it aims to quantify the relationship between social determinants of health, which are linked to epigenetic outcomes, and cancer incidence and mortality.

Epigenetics refers to reversible, non-structural changes to an individual's DNA that are due to behavioral and environmental health. These changes occur through either DNA methylation or histone modification, with both breast and colorectal cancer having observed epigenetic DNA methylation patterns [3]. In observing these patterns, several studies have shown that increases in breast, colorectal, and lung cancer incidence and mortality are linked to poor social determinants of health [4], [5], [6], [7].

Social determinants of health (SDOH) are the non-medical and cumulative factors that influence an individual's health outcomes. They represent living conditions, working environments, socioeconomic status, community belonging, healthcare access, and education [8].

Existing studies have demonstrated associations between individual-level SDOH factors and cancer outcomes, but county-level quantification of these relationships in North Carolina remains limited. This gap limits the precision of state-level policy design and highlights the need for an integrated computational-public health approach. This analysis is critical for state policymakers seeking to propose targeted interventions and broader public health measures.

This study quantifies the relationship between SDOH and cancer outcomes for individuals in North Carolina, using data from the NC Department of Health and Human Services and NC Institute of Medicine. It examines whether SDOH effects are cancer-specific or generalizable, identifies key factors associated with outcomes, discusses policy implications, and data accessibility barriers that constrain population-level epigenetic research.

II. METHODOLOGY

A. Study Design

This cross-sectional observational study examples relationships between SDOH and cancer outcomes across all 100 North Carolina counties using SDOH data from 2021, and aggregate data on cancer incidence and mortality from 2019-2023. This timeline allows for the observed SDOH variables to take effect, ensuring that the dependent variables in the analysis are in response to the snapshot independent variables.

The unit of analysis is the county. Dependent variables include incidence and mortality for breast, lung, and colorectal cancers. These cancers are evaluated both independently and generally, and were selected from the available data (breast, colorectal, lung, melanoma, cervical, prostate) due to their

This paper was submitted to Dr. Minh Tang as a course assignment for CSC/ST 442: Introduction to Data Science on December 9, 2025.

Madison E. Thompson is with the Department of Computer Science, Department of Integrative Humanities and Social Sciences, Department of Biological Sciences, Department of Mathematics, Department of Molecular Biomedical Sciences, North Carolina State University, Raleigh, NC 27695 USA (e-mail: methom22@ncsu.edu).

established relationship with epigenetic factors, which are linked to SDOH. This ensures that the analysis is relevant.

Independent variables are SDOH metrics, including insurance status, access to primary care providers (PCPs), college graduation, adult smoking, poverty, food insecurity, transportation, and air pollution. These variables each contribute to cancer-outcome models and the aggregate data.

This study addresses three research questions: (1) Which SDOH factors show strongest associations with cancer outcomes? (2) Do SDOH effects generalize across cancer types or are relationships cancer-specific? (3) What policy-relevant insights emerge from county-level SDOH-cancer associations?

B. Data Sources and Collection

The 2019-2023 aggregate cancer incidence and mortality data used in this study were obtained from the North Carolina Department of Health and Human Services, NC State Center for Health Statistics (SCHS) [9], [10]. County-level SDOH metrics were drawn from the North Carolina Institute of Medicine (NCIOM) 2021 statewide SDOH snapshot [11].

The SCHS data were provided in PDF format and required manual extraction and conversion into CSV files prior to analysis. The NCIOM dataset was provided as an Excel file and likewise converted into a CSV format for consistency in analysis. For all sources, blank rows were removed and counties were excluded from analysis for a cancer, as well as for the general data when complete data were not available.

Before selecting the sources above, the author attempted to use several national databases, primarily those of genomic consequence to connect epigenetic indicators into the analysis.

The National Cancer Institute Genomic Data Commons (GDC) provides useful information on the genotypes of various cancers, including DNA methylation arrays for breast, colorectal, and lung cancer [12]. However, this data does not include geographic information or metadata that could be used to draw linkage between policy and cancer incidence.

To address this gap, the author then gained access to the National Institute of Health *All of Us* database, which includes case-level information for demographics, cancer incidence, mortality, and genetic presentation of cancers. However, usage of this database requires extensive training and reporting, makes research workspaces publicly available, and has vague restrictions on how data may be utilized. Further, there are restrictions on importing other datasets, which would make the cross-sectional analysis the author was attempting to perform infeasible [13]. An inability to utilize this dataset made it infeasible to directly link epigenetic trends to cancer outcomes in any area.

Considering a macro-level analysis, the author also attempted to gain access to the NCI Surveillance, Epidemiology, and End Results Program (SEER) database, but was blocked by a requirement of institutional registration. As a result, the author was unable to look at data across multiple decades, or across multiple states [14].

Collectively, these limitations resulted in the author deciding to focus on a more in-depth analysis for the state and counties of North Carolina, as this data is publicly available and presented on a timescale that is appropriate for epidemiological evaluation. While this decision may not be ideal, it serves as a proof-of-concept and preliminary study for future works that evaluate the relationship between SDOH and cancer outcomes on a national or international level. Further, the limited scope protects the internal validity of the results presented and allows for a more thorough exploration of the data available.

A strong future direction, with more time and resources, would be to find the SDOH, cancer incidence, and cancer mortality data for each state, and see if there is a national correlation. However, this also introduces many confounding variables on the policy level and could weaken interpretations of how the data might be utilized for social or political action.

C. Variable Operationalization

This study operationalized cancer outcomes using age-adjusted incidence and mortality rates per 100,000 individuals as reported by SCHS, and county-level SDOH predictors as reported by NCIOM. Due to SDOH metrics being reported heterogeneously (percentages, counts, index), all predictors were standardized using z-scores prior to modeling to ensure comparability across variables and cancer types. Tables 1 and 2 summarize each variable included in the analysis. Table I contains the source, model role, and definition for each variable, and Table II contains the unit and transformation.

TABLE I
VARIABLE OPERATIONALIZATION: DEFINITIONS

	Variable Name	Source	Definition
Indicators	Breast cancer incidence	SCHS	Age-adjusted incidence rate per 100,000 individuals (2019-2023)
	Breast cancer mortality	SCHS	Age-adjusted mortality rate per 100,000 individuals (2019-2023)
	Colorectal cancer incidence	SCHS	Age-adjusted incidence rate per 100,000 individuals (2019-2023)
	Colorectal cancer mortality	SCHS	Age-adjusted mortality rate per 100,000 individuals (2019-2023)
	Lung cancer incidence	SCHS	Age-adjusted incidence rate per 100,000 individuals (2019-2023)
	Lung cancer mortality	SCHS	Age-adjusted mortality rate per 100,000 individuals (2019-2023)
Predictors	Uninsured adults	NCIOM	Percentage of uninsured adults aged 18-64 (2018)
	PCP access	NCIOM	Number of PCPs per 10,000 individuals (2019)
	College graduation	NCIOM	Percentage of adults aged 25+ with a Bachelor's degree or higher (2013-2017)
	Adult smoking	NCIOM	Percentage of adults who are current smokers (2017)
	Poverty	NCIOM	Percentage of population living in poverty (2019)
	Food insecurity	NCIOM	Percentage of population that is food insecure (2018)
	Transportation	NCIOM	Percentage of households without access to a vehicle (2017)
	Air pollution	NCIOM	Average daily density of fine particulate matter (mcg per cubic meter) (2014)

TABLE II
VARIABLE OPERATIONALIZATION: TRANSFORMATIONS

	Variable Name	Unit	Transformation
Indicators	Breast cancer incidence	Per 100,000	None
	Breast cancer mortality	Per 100,000	None
	Colorectal cancer incidence	Per 100,000	None
	Colorectal cancer mortality	Per 100,000	None
	Lung cancer incidence	Per 100,000	None
	Lung cancer mortality	Per 100,000	None
Predictors	Uninsured adults	Percentage	z-score
	PCP access	Per 10,000	z-score
	College graduation	Percentage	z-score
	Adult smoking	Percentage	z-score
	Poverty	Percentage	z-score
	Food insecurity	Percentage	z-score
	Transportation	Percentage	z-score
	Air pollution	Index	z-score

Cancer outcomes retain their original units to preserve interpretability after model evaluation. All SDOH predictors were standardized using z-scores to create a comparable scale for regression analysis.

D. Exploratory Data Analysis

Exploratory data analysis (EDA) was conducted to characterize the distributions of cancer outcomes and SDOH predictors across North Carolina counties, and to identify potential patterns and outliers prior to modeling. Summary statistics were computed for all variables (Table III), and boxplots (Figure 1) and a pairwise correlation matrix (Figure 2) were developed for visual inspection.

TABLE III
SUMMARY STATISTICS FOR SELECT SDOH PREDICTORS

Predictor	Mean	Median	SD	Min	Max
Adult smoking	-4.7184e-17	0	1	-2.7396	3.2378
Air pollution	-3.3085e-16	0.1356	1	-2.1050	1.8159
College graduation	-6.5226e-17	-0.2054	1	-1.474	3.8107
Food insecurity	-1.0492e-16	-0.1426	1	-2.3210	2.3646
PCPs	-1.5543e-17	-0.1194	1	-1.7182	5.4925
Population	-8.8817e-17	-0.2890	1	-0.5931	5.9209
Poverty	-1.5543e-16	-0.1265	1	-1.7890	3.3038
Uninsured adults	1.8763e-16	-0.2220	1	-1.8310	3.1390
Transportation	3.844e-17	-0.1698	1	-1.8549	2.6245

Table III verifies the z-score standardization of the data, demonstrating a z-score mean ≈ 0 and SD = 1 for each metric. This certifies that the selected predictors may be used to perform transformations on the indicator variables.

The range of standardized values reveals heterogeneity across counties, with PCPs having a range of (-1.7182, 5.4925), poverty having a range of (-1.7890, 3.3038).

Median values reveal some distributional skew, which is also well-visualized in Figure 1. Most significant are differences in educational attainment (college graduation) (median = -0.2054) and population (median = -0.2890).

It should be noted that population itself is not a SDOH variable, but that it is salient to public policy analysis, where it will be utilized to inform discussion.

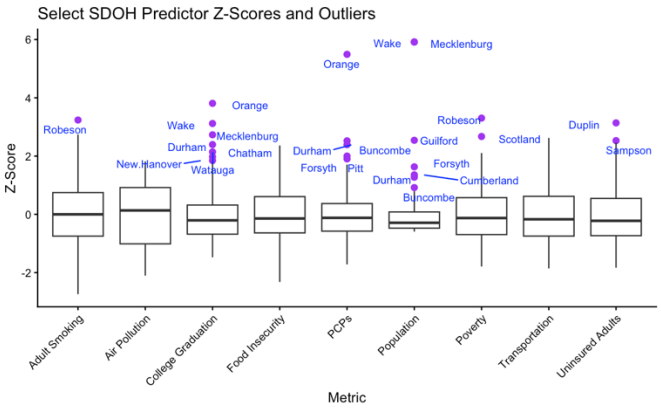

Fig. 1. Box plots of standardized SDOH predictors across NC counties. Notable outliers include Wake and Mecklenburg (high education, population) and Robeson (high smoking prevalence and lowest life expectancy).

Figure 1 identifies several outliers in the data, most notably Wake and Mecklenburg having the highest number of college graduates and largest populations, indicating a relationship between urbanization and higher education. On the flip side, Robeson County, which has the lowest life expectancy of any county in NC, has the highest rate for adult smoking. Interestingly, PCPs are not isolated to urban areas.

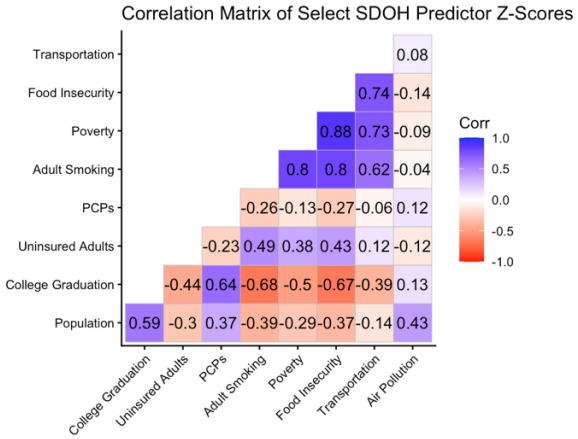

Fig. 2. Pearson correlation matrix of SDOH predictors. Strong positive correlations exist between poverty, food insecurity, and smoking (r > 0.80), suggesting a clustered socioeconomic vulnerability profile.

Figure 2 presents interesting data, showing a decrease in almost every risk factor when correlated with higher education,

with adult smoking and food insecurity showing the greatest inversely proportionate relationship. College graduates also seem to populate areas with PCPs (0.64 Pearson correlation). On the converse, it can be seen with significant statistical power that food insecurity and poverty are linked (0.88), as well as poverty and adult smoking (0.8), and food insecurity and adult smoking (0.8). These clusters are followed by weaker, but similarly compelling evidence for a relationship between transportation and poverty (0.73) and transportation and food insecurity (0.74).

These findings suggest a clustered socioeconomic vulnerability profile where poverty, food insecurity, smoking, and transportation barriers co-occur. This multicollinearity will require careful handling in regression models (see Section II.G). The protective association of educational attainment across multiple risk factors suggests it may serve as a key moderating variable in cancer-specific analyses.

E. Regression Modeling: General Cross-Cancer Analysis

Four pooled models were constructed to identify generalizable SDOH associations with cancer outcomes. Two models evaluate the pooled incidence rates across all three cancer types, and two evaluate the pooled mortality rates. Given the clustering shown in Figure 2, these models need to account for multicollinearity, so both Ordinary Least Squares (OLS) models and Ridge regression were fit for each outcome. The pooled model specification is:

$$Y_i = \beta_0 + \beta_1(Uninsured\ Adults)_i + \beta_2(PCP)_i + \beta_3(Education)_i + \beta_4(Adult\ Smoking)_i + \beta_5(Poverty)_i + \beta_6(Food\ Insecurity)_i + \beta_7(Transportation)_i + \beta_8(Air\ Pollution)_i + \varepsilon_i \quad (1)$$

Where Y_i represents age-adjusted cancer outcome (incidence or mortality) per 100,000 individuals for observation i, all represents predictors are z-scored, and ε_i is the error term.

The Ridge regression model uses the same specification but adds an L2 penalty term:

$$\text{minimize: } \Sigma(Y_i + \hat{Y}_i)^2 + \lambda\Sigma{\beta_j}^2 \quad (2)$$

Where λ is the regularization parameter selected via 10-fold cross-validation to minimize mean squared error (MSE). Ridge shrinks coefficient estimates toward zero proportionality, which accounts for the multicollinearity seen in the dataset without removing any predictors.

Ridge regression was selected to address multicollinearity concerns while retaining predictors in the model. Spline and LASSO were avoided because Spline regression assumes a lack of linearity, but the data demonstrates linear connection between multiple SDOH predictors. LASSO is utilized later in the study, and a diverse set of data analysis was desired. With Ridge, no predictors are removed from the model, and independently selected regulation parameter, λ, uses 10-fold cross-validation to minimize mean squared prediction error. This is a data-driven approach that ensures penalty strength is appropriate without compromising full exploration of the data, which is essential to this study, as it is seeking to determine which parameters hold most significance.

For each pooled model, R^2, adjusted R^2, and AIC are reported to assess overall fit. OLS and Ridge coefficient estimates are compared to evaluate the significance of the observed multicollinearity and whether the correlation is artificially inflating predictor variables in the OLS model.

Comparison with a preference for the variables supported by Ridge regression is conducted to determine which SDOH predictors are significant predictors for cancer outcomes.

Comparison between the predictors suggested by incidence and mortality models will inform which SDOH predictors are significant for prevention, and which hold more weight for treatment.

Coefficients are interpreted as standardized effect sizes, representing the change in chancer rate per one standard deviation increase in each SDOH predictor.

F. Regression Modeling: Cancer-Outcome Analysis

Evaluating pooled incidence and mortality is useful for understanding with SDOH predicators may be generalizable to all cancers, but further analysis can determine whether certain SDOH factors carry varying weight for each cancer-outcome combination (n=6, breast cancer incidence, breast cancer mortality, lung cancer incidence, lung cancer mortality, colorectal cancer incidence, colorectal cancer mortality). To determine this information, LASSO regression models were fitted for each outcome.

LASSO regression models perform automatic variable selection by shrinking some coefficients to exactly zero, identifying which SDOH parameters are most strongly associated with each cancer-outcome combination. This approach allows predictors to be selected via a model, avoiding researcher bias and addressing multicollinearity through L1 penalization.

Each model uses the LASSO objective function:

$$\text{minimize: } \Sigma(Y_i + \hat{Y}_i)^2 + \lambda\Sigma|\beta_j| \quad (3)$$

Where Y_i represents age-adjusted cancer outcome (incidence or mortality) per 100,000 individuals for county i, all predictors are z-scored, and λ is the L1 penalty parameter.

The full model specification before variable selection is equivalent to (1), where λ was selected independently via 10-fold cross-validation to minimize MSE. Unlike the minimization function in (2), some predictors are not retained after penalization, meaning (1) may function fundamentally differently for each of the 6 LASSO models. Any variable retained with a non-zero coefficient after penalization is regarded as a significant predictor for that cancer outcome.

Separating incidence and mortality models for each cancer allows for a deeper exploration of the relationship between SDOH predictors and cancer outcome indicators. This allows for the identification of pathways that may not be observed in the generalized data, and shows differences between epigenetic determinants, which all cancers are known to be modulated by.

For each cancer-specific model, R^2 and the cross-validated mean square error (CV-MSE) are reported to assess overall model fit. The SDOH variables that are retained, and in what magnitude, for each model are also reported, as well as those that are eliminated.

Patterns are compared across cancer types, addressing whether there exists SDOH factors that are universally important, other factors that are cancer-specific relationships, and whether incidence and mortality predictors vary within cancer types (is there a difference in which predictors are significant for incidence versus mortality).

G. Model Validation and Diagnostics

Model performance was assessed using multiple qualitative and quantitative measures. For OLS and Ridge regression (generalized models), R^2, adjusted R^2, and AIC were used to measure the proportion of variance explained, as well as to assess overall model fit. AIC provides a comparative measure accounting for both fit and parsimony, where lower values are associated with better balance.

Each LASSO model uses CV-MSE as the primary fit measure, as the λ selection process inherently balances prediction accuracy and model complexity. R^2 was also reported for LASSO models, while adjusted R^2 was not used because the effective number of parameters varies depending on whether some predictors are minimized to zero.

For OLS models in Section II.E, standard regression diagnostics were conducted to verify assumptions.

- Residual plots were examined for systematic patterns indication violations
- Q-Q plots assessed whether residuals follow a normal distribution, validating inference procedures
- Variance Inflation Factors (VIF) were calculated for each predictor, where VIF > 5 indicates problematic multicollinearity
- Cook's distance (D) was calculated to identify observations with disproportionate influence on coefficient estimated. Counties with $D > \frac{4}{n}$ were flagged as potentially influential

These diagnostics justify the use of Ridge and LASSO and identify which assumptions are violated in standard OLS.

Ridge and LASSO models used 10-fold cross-validation to select penalty parameters. The dataset was randomly partitioned into 10 folds, with each fold being once held out as a validation set while the other 9 were used for training data. The partition that minimized MSE was selected for each model, and the process was independently repeated for each model, ensuring optimized penalty strength for each cancer-outcome combination without overfitting.

Finally, the exploratory findings were utilized to ensure robustness of findings. Data identified as outliers, as well as data flagged by Cook's distance, was removed and models were refitted. If the refitted models remained relatively stable, it indicates that they are relatively robust and that outliers do not create substantial change. If outlier data points carry significant weight in fitting, it indicates that the relationship between select SDOH predictors and cancer outcome indicators may not be as present as the model indicates.

H. Policy Context Analysis

Understanding how cancer outcomes are correlated to SDOH predictors in all 100 North Carolina counties is critical to determining future policy directions, especially as it relates to decisions about both preventative care and treatment accessibility. The data collected in this study is utilized to briefly analyze the effectiveness of existing public health policy in North Carolina, and to suggest reforms where they may be necessary. Some topics that are discussed include the 2023 Medicaid expansion, NC SNAP outreach efforts, and the diversity of county health programs in the state.

III. Results

A. Pooled Analysis: Incidence

Table IV presents coefficient estimates for OLS and Ridge regression models predicting pooled cancer incidence across the types of interest. The OLS model was a poor fit ($R^2 = 0.022$, adjusted $R^2 = -0.006$, AIC = 3316) and was not statistically significant ($F(8,283) = 0.78$, $p = 0.62$). No individual SDOH predictors reached statistical significance, with coefficient estimates ranging from -8.83 (College Graduation) to 4.26 (Air Pollution).

Ridge regression with cross-validated λ selection yield unexpected results, producing an even poorer fit ($R^2 = 0.010$, adjusted $R^2 = -0.018$, AIC = 2486). Ridge coefficients were heavily minimized toward zero, but predictive performance was not improved. This indicates that it is unlikely that any of the selected SDOH predictors play a significant role in determining cancer outcomes, and that multicollinearity is not a primary limited for this dataset.

TABLE IV
OLS and Ridge Statistics for Pooled Cancer Incidence Rates

Predictor	OLS Est.	SE	p	Ridge Est.
Adult smoking	0.05	8.49	0.995	0.67
Air pollution	4.26	4.39	0.332	0.42
College graduation	-8.83	8.41	0.294	-0.83
Food insecurity	1.41	11.53	0.903	0.69
PCPs	2.17	5.70	0.704	-0.28
Poverty	1.52	10.14	0.881	0.57
Uninsured adults	-0.81	4.99	0.871	0.23
Transportation	0.85	7.04	0.904	0.58
Model Fit				
R^2	0.022			0.010
Adj R^2	-0.006			-0.018
AIC	3326			2487

The poor performance of the pooled incidence model suggests that aggregating across cancer types obscures cancer-specific SDOH relationships. The failure of the Ridge regularization to improve fit confirms that multicollinearity is not the primary limitation of this approach, which is surprising

given the EDA. This data does serve to demonstrate the necessity of the analyses in Section III.C and III.D, as a cancer-specific breakdown can indicate whether SDOH is simply ineffective when pooled, or if the selected metrics simply have weak predictive power.

B. Pooled Analysis: Mortality

Table V presents the parallel analysis for pooled mortality rates. The OLS model showed improved fit compared to incidence ($R^2 = 0.072$, adjusted $R^2 = 0.039$, AIC = 1827), and minor statistical significance ($F(8,226) = 2.19$, $p = 0.029$). College graduation showed protective statistical significance ($\beta = -3.24, p = 0.066$) at the 0.1 level, as anticipated from the EDA.

Ridge regression with cross-validated λ selected yielded a similar fit ($R^2 = 0.058$, adjusted $R^2 = 0.025$, AIC = 1213). The protective college graduation coefficient was substantially shrunk (OLS = -3.24, Ridge = -1.01), indicating some sensitivity to multicollinearity despite the protective predictor.

TABLE V
OLS AND RIDGE STATISTICS FOR POOLED CANCER MORTALITY RATES

Predictor	OLS Est.	SE	p	Ridge Est.
Adult smoking	0.57	1.81	0.751	0.57
Air pollution	-0.27	0.93	0.771	-0.17
College graduation	-3.24	1.76	0.066	-1.01
Food insecurity	1.11	2.47	0.652	0.54
PCPs	0.57	1.22	0.639	-0.32
Poverty	0.12	2.15	0.955	0.31
Uninsured adults	-0.59	1.09	0.589	0.08
Transportation	-0.74	1.52	0.628	0.15
Model Fit				
R^2	0.072			0.058
Adj R^2	0.039			0.025
AIC	1827			1213

The statistical significance of higher educational attainment suggests that it is associated with lower cancer mortality across cancer types at the county level. However, only this protective predictor reached a significance threshold, suggesting that the original assumptions surrounding multicollinearity may have been misguided. This presents an interesting perspective on how protective SDOH predictors may be more relevant than those assumed to worsen health outcomes.

The Ridge regression model suggests that the lag data between the 2021 snapshot (which has data from 2013-2019) may be a better predictor of mortality, which would make sociological sense. Once a SDOH factor has been recorded, it will take time for the systemic effects to appear in population health data.

C. Cancer-Outcome LASSO Modeling and Analysis

A total of six models were trained, one for each cancer indicator (type and outcome) The models fit are summarized in Table VI. In contrast to pooled models, the disaggregated models demonstrated sustainably improved explanatory power. In contrast to the pooled models, the cancer-outcome incidence models outperformed the mortality models for lung ($R^2 = 0.489$ v. $R^2 = 0.469$) and colorectal cancers ($R^2 = 0.496$ v. $R^2 = 0.242$). Breast cancer was anticipated to be one of the more predictable outcomes, as it is well-studied in epigenetics, however, it showed the least correlation between SDOH parameters and both incidence and mortality outcomes, with mortality prediction at a weak 8.8% ($R^2 = 0.394$ v. $R^2 = 0.088$). Further, LASSO only selected one predictor (Food Insecurity) for breast cancer mortality, but retained all 8 predictors colorectal cancer incidence, which had the highest R^2 value.

TABLE VI
LASSO MODEL PERFORMANCE SUMMARY

Cancer Outcome	R^2	CV-MSE	λ	Variables Retained
Breast Inc.	0.394	379	0.358	6
Breast Mor.	0.088	17.5	1.100	1
Lung Inc.	0.489	83.6	0.133	7
Lung Mor.	0.469	31.4	0.092	7
Colorectal Inc.	0.496	31.9	0.142	8
Colorectal Mor.	0.242	8.16	0.327	5

In the above table, R^2 is calculated from CV-MSE. All models use 10-fold cross-validation with λ selected to minimize prediction error.

When considering the larger context of this work, it is critical that the actual impact of SDOH predictors is determined. To determine which predictors have a significant impact on either incidence or mortality and may warrant discussion for public health professionals in North Carolina, Table VII was constructed. Table VII displays which SDOH predictors were retained versus eliminated by LASSO for each cancer-outcome combination.

TABLE VII
LASSO MODEL CANCER-OUTCOME COEFFICIENTS

Predictor	BI	BM	LI	LM	CI	CM
Adult smoking	2.43	-	0.83	2.33	1.10	0.12
Air pollution	10.97	-	1.55	1.68	0.54	-0.05
College graduation	-9.62	-	-8.34	-5.40	-3.83	-0.89
Food insecurity	-	1.06	2.06	-	0.21	0.57
PCPs	3.38	-	0.86	0.86	0.03	-
Poverty	0	-	-	-0.58	0.99	0.12
Uninsured adults	0.35	-	-1.90	-1.53	-0.90	-
Transportation	4.70	-	0.24	-0.65	1.33	-

In the above table, abbreviations were used for cancer outcomes, with BI represents breast incidence, BM breast mortality, and so forth. A - indicates that the predictor was eliminated by the model.

Table VII shows that college graduation is the most consistent predictor, retained in five of six models. In the incidence models, a stronger retention was observed (6-8 variables), compared to less so for mortality (wide range of 1-7). The outlier data, breast cancer mortality, retained only one

parameter and had the lowest predictive power. This warrants further study, as other factors not observed in this paper are likely forming the mortality pathway for breast cancer patients in North Carolina.

It is also of note that uninsured adults is a protective predictor for half of the models, which is unexpected. It is anticipated that a lack of health insurance would be correlated with an increase in cancer incidence or mentality, and this may warrant future study.

Figure 3 visualizes the data in Table III through vectorization. It plots incidence and mortality parameters by standardized coefficient for each SDOH, showing trends and within across cancer outcomes.

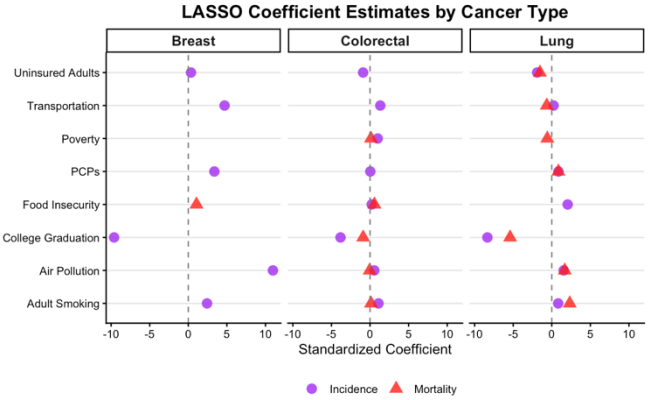

Fig. 3. Vectorized cancer-specific SDOH coefficient visualization, where purple circles represent incidence rates and red triangles represent mortality.

Notably, college graduation shows consistent negative coefficients (protective), and adult smoking shows more positive coefficients than other cancer types, as expected given the EDA. Interestingly, colorectal cancers remain around zero, despite having the highest predictive power and retaining the most predictors. The data for breast cancer is considerably scattered, showing how a poorly fit model may present. Lung cancer shows similar magnitudes for incidence and mortality rates, indicating that the two may be related.

D. Model Evaluation and Robustness

As outlined in Section II.G, residual and Q-Q plots, Cook's distance, and VIFs were calculated to evaluate the validity and robustness of the models presented.

VIFs were computed for all predictors in the pooled OLS models, which was essential given the anticipated multilinearity from Figure 2. Most predictors fell below VIF < 5, however, poverty (VIF = 6.03) and Food Insecurity (VIF = 7.72) exceeded the provided threshold. Given that these SDOH predictors were both members in the cluster found during EDA, this is unsurprising and justify the author's decision to use Ridge regression to stabilize coefficient estimates. Because neither value exceeds VIF > 10, they are still suitable to use for modeling with stabilization.

Figure 4 shows the standard OLS diagnostic plots (Residuals vs Fitted, Q-Q plot, Scale-Location, Residuals vs Leverage), which demonstrate multiple violations of linear regression assumptions.

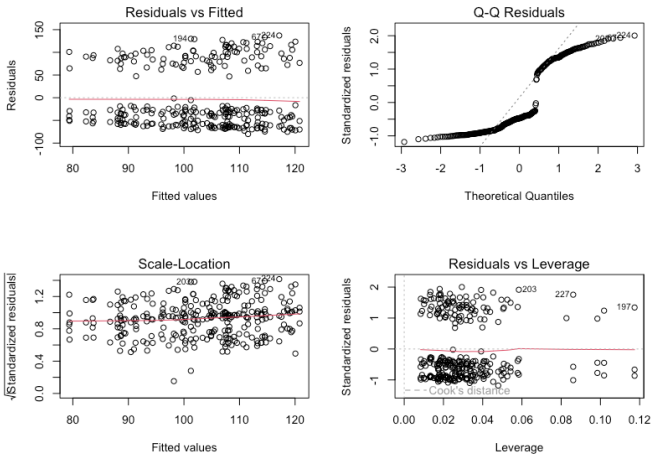

Fig. 3. Standard OLS diagnostic plots, indicating violations of linear regression assumptions that justify Ridge regression and LASSO modeling.

As seen above, the Residuals vs Fitted plot shows structured, non-random residual patterns, suggesting nonlinearity in the data. The Q-Q plot shows substantial departures from normality, which was confirmed by a Shapiro-Wilk test ($W = 0.81, p < 2.2\text{e-}16$) that rejected normality of residuals. The Scale-Location plot shows increasing variance at higher fitted values, suggesting inconsistent variance across the dataset. The Residuals vs Leverage plot also shows that some counties have significant pull on the data, which indicates that a pooled linear model would be structurally insufficient. All this information further justifies the decision to use Ridge regression and later cancer-outcome specific LASSO analysis.

Cook's distance analysis identified nine influential observations exceeding the threshold of $D > \frac{4}{n}$, namely Wake, Robeson, Durham, Dublin, Onslow, Orange, Scotland, Swain, and Pasquotank. Wake county is densely population and has many higher education institutions, which likely makes it an outlier for many datapoints. The other counties of influence represent rural areas, with Robeson County having the greatest influence at ($D = 0.032$). This county is considered to have the worst SDOH outcomes of any county in NC, so it is an expected outlier [11].

For LASSO models, cross-validation inherently provides validation through held-out fold predictions. The substantial improvement in model predications when shifting from pooled to disaggregated (cancer-outcome) demonstrates the impact of the above metrics. The pooled model is weak due to inappropriate aggregation, rather than from data quality issues or outlier influence.

IV. Discussion

A. Predictive Performance Across Models

Predictive modeling showed varying levels of success, with the pooled OLS data performing with no substantial impact ($R^2 = 0.022$). This is likely due to the assumption violations demonstrated in Section III.D, but it can be concluded that

SDOH predictors cannot be used to determine pooled cancer outcomes using the data provided.

This interpretation is further justified by the poorly performing Ridge regression models, with the pooled incidence model having the lowest impact of any model analyzed ($R^2 = 0.010$). For mortality, Ridge showed mild improvement ($R^2 = 0.072$) but still did not cross the minimum significance established by cancer-outcome specific models ($R^2 = 0.088$). This suggests that pooled data, at least from this source, simply has poor predictive power. Given that Ridge regression handles multicollinearity, and no predictors have a VIF > 10, it can be assumed that pooled models are simply not linked to SDOH for this data source.

LASSO modeling for cancer-outcome analysis provided the most useful data, demonstrating that there exists a substantial correlation between some SDOH predictors and various cancer outcomes. These models outperformed the author's expectations, with lung cancer incidence ($R^2 = 0.489$), lung cancer mortality ($R^2 = 0.469$), and colorectal cancer incidence ($R^2 = 0.496$) being predicted by SDOH variables alone 47%-50% of the time. These values show that the data is not flawed, but that SDOH predictors likely have targeted effects on various outcomes, meaning they will impact each cancer diagnosis and prognosis differently. This presents important information for public policy officials as they look to engage with preventative and palliative care models for residents of North Carolina.

B. Predictor Retention Interpretations and Impact

LASSO analysis retained some key predictors across cancer outcomes, indicating which are significant to each. Education was a protective predictor across all cancer outcomes, meaning education seems to be inversely proportionate to poor health outcomes, particularly those facilitated by SDOH predictors. Sociologically, this makes sense, as higher education has long been linked to positive health outcomes for individuals and their households [15]. Figure 3 provides the most useful insight into the significance of predictors.

First, breast cancer incidence and mortality are seemingly randomly spread, and have the lowest model power out of any cancer-outcome group ($R^2 = 0.394, R^2 = 0.088$). This indicates that breast cancer, which has well-established epigenetic linkages, is likely being modulated by some external factor that the model attempted to account for. Given the radical variance between parameters, I hypothesis there exists significant differences in cancer outcomes in rural and urban areas, and that the shifts shown are representing the difference in county conditions, but not the actual predictors that modulate the outcomes.

Colorectal cancer had the greatest predictor power and is the closest to center. This is an interesting turn, as one may expect it to have distinct indicators across the SDOH predictors. The retention to center suggests that the SDOH variables likely have a limited effect, suggesting that the LASSO model is accounting for diffused signal rather than SDOH as a driving factor in cancer outcomes. However, the predictive power is still strong, and if anything, this warrants future study to determine what other variables this model may be combined with to increase predictive power.

Lung cancer follows a similar pattern to colorectal cancer, with shifts that give more predictive power to air pollution and smoking. This makes logical sense, as both SDOH predictors are known to negatively affect lung health.

Access to a PCP appears to show an increase in cancer incidence, but this is likely due to screening and detection, versus actual increased diagnosis. Further, there is no significant difference in mortality, indicating that cancer outcomes are not different with access to a PCP (which is somewhat alarming), supporting the hypothesis that it is an artificially inflated variable related to screening, and should be removed in future analysis.

C. Incidence and Mortality Differences

Before discussing policy, it is worth noting the differences between incidence and mortality predictors. Incidence predictors are used to determine whether someone may be exposed to risk factors that lead to the development of certain cancers, such as smoking, air pollution, and food insecurity. Mortality predictors reflect systemic disadvantage, such as a lack of insurance, lack of PCP access, and lack of transportation to and from healthcare centers.

These factors appear to affect each model differently, indicating that incidence and mortality are in fact distinct. This is best seen in the pooled mortality data, where poor SDOH has some substantial interpretations, unlike incidence. This may indicate that mortality is more significant to SDOH predictors than incidence, which would be sociologically valid. If someone is already suffering from poor health, it makes every additional condition exponentially more exacerbating.

Interestingly, incidence and mortality are often distinct, with predictors failing to overlap in coefficient outputs (Figure 3). This indicates that future study may show patterns where incidence is caused by certain SDOH predictors, and mortality is further mediated by a different subset of predictors. This could point researchers to new social pipelines and help public health professionals better understand how to provide care for those in need.

D. Implications for Policymakers

Policymakers maintain a need to be continually aware of changing environments, as well as ways to optimize care for their constituents. This data provides multiple implications for policymakers, namely: higher education matters, screening and access to PCPs is likely to lead to earlier diagnosis, and poverty worsens lives. More extremely, education saves lives, and poverty kills.

The data presented in this report clearly demonstrates that even when every other factor may be underdetermined, higher education is consistently correlated with positive health outcomes. Receiving an education makes individuals more likely to be aware of screenings, seek out PCPs, and receive adequate medical care before their conditions are no longer treatable. The PCP predictor demonstrates that access to care and screening can lead to diagnosis, and those whom complete higher education are more likely to seek out such care [15].

On the converse, significant multicollinearity is observed between key SDOH variables such as food insecurity, poverty, transportation, and smoking. Despite not being of great significance in this study, these SDOH predictors were

identified by public health professionals far more qualified than the author, and for good reason. They are known to affect the health of individuals and communities, and addressing poverty is a critical step toward health equity for residents of North Carolina.

E. Limitations

A major limitation of this project is the data available. As discussed in the methodology, the author made many attempts to access large-scale databases that contain useful biological information. This information could be utilized to verify whether cancer incidences are epigenetic and filter out noisy data that likely weakens correlations. The restriction of these databases has some merit, admittedly protecting patient anonymity and preventing harmful misuse to further propagate stereotypes or eugenic ideals. However, making this data inaccessible to researchers also greatly limits their ability to identify key problems in public health and begin work to address these issues with interdisciplinary teams. If public health research is to meaningful expand, it must keep up with technology.

The data that was able to be collected is county-level SDOH data, county-level cancer incidence, and county-level cancer mortality. This data has several flaws, one most notably being the timeline. The 2021 snapshot provided by the NCIOM is actually data from 2014-2019, based on what they were able to pull together at the time. This means that there is not a set ability to say, "5 years after X predictor reaches Y threshold, A change in B variable may occur," which this analysis could have lended itself to producing. This lag is not a major determent, as it at least allows SDOH predictors time to affect a population, but standardized data would be significant.

It should also be noted that New Hanover County was excluded, as it failed to provide enough data to make the NC SCHS report. A lack of data is a definite concern with this report. While many variables are considered, there exists only one datapoint per variable. A report over 20 years would be more meaningful.

It should also be noted that SDOH variables were selected at the author's discretion. This decision was made based on which variables were likely associated with adult SDOH outcomes and which best aligned with the parameters set by the NIH. However, there could easily be a variable in the data that was not analyzed that has greater significance than any variable selected.

Lastly, there are clearly confounding variables not represented. Breast cancer diagnosis and treatment is one of the major contributors to epigenetics research and knowing that it is showing the least statistical significance indicates that some other variable, SDOH or otherwise, must be playing a major effect in epigenetic cancers. Identifying and analyzing these data points is a critical next step for any author looking to continue this work.

V. Conclusion

As the world continues to move into a big data landscape, it is increasingly imperative that researchers holistically evaluate healthcare data and outcomes. Preliminary studies such as the one presented in this analysis provide a foundation for dialogue between engineers, life scientists, and policy makers. The author firmly believes that the development and utilization of interdisciplinary teams is critical to performing meaningful research in the information age.

In this study, the relationships between select SDOH predictors and cancer outcomes (incidence and mortality) for breast, lung, and colorectal cancers is evaluated for all 100 counties in North Carolina. The findings indicate that college graduation is a protective parameter for public health in North Carolina. They also indicate that poverty, food insecurity, transportation, and adult smoking are involved in a complex multilinear relationship that may worsen health outcomes for those of a lower socioeconomic status.

It was found that evaluating the relationship between SDOH predictors and cancer outcomes has some success, with LASSO modeling predicting lung cancer incidence ($R^2 = 0.489$), lung cancer mortality ($R^2 = 0.469$), and colorectal cancer incidence ($R^2 = 0.496$) at rates beyond the author's expectations.

In the future, this research could be built upon to co-evaluate SDOH and biological predictors, or to inform how SDOH predictors should be addressed in public health policy.

VI. Supplementary Material

The source code used for this report can be found at the following repository:

https://github.com/madisonthompson27/nc-county-epigenetics/

VII. References

[1] CDC, "FastStats," Leading Causes of Death. Accessed: Dec. 08, 2025. [Online]. Available: https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

[2] NCDHHS, "Leading Causes of Death," Leading Causes of Death in North Carolina 2019. Accessed: Dec. 08, 2025. [Online]. Available: https://schs.dph.ncdhhs.gov/interactive/query/lcd/getleadcauses.cfm

[3] CDC, "Epigenetics, Health, and Disease," Genomics and Your Health. Accessed: Dec. 08, 2025. [Online]. Available: https://www.cdc.gov/genomics-and-health/epigenetics/index.html

[4] A. D. Williams and T.-A. Moo, "The Impact of Socioeconomic Status and Social Determinants of Health on Disparities in Breast Cancer Incidence, Treatment, and Outcomes," *Curr. Breast Cancer Rep.*, vol. 15, no. 1, pp. 30–36, Mar. 2023, doi: 10.1007/s12609-023-00473-7.

[5] M. K. Lorentsen and H. K. Sanoff, "Social Determinants of Health and the Link to Colorectal Cancer Outcomes," *Curr. Treat. Options Oncol.*, vol. 25, no. 4, pp. 453–464, Apr. 2024, doi: 10.1007/s11864-024-01191-7.

[6] D. Shin, M. D. C. Fishman, M. Ngo, J. Wang, and C. A. LeBedis, "The Impact of Social Determinants of Health on Lung Cancer Screening Utilization," *J. Am. Coll.*

Radiol., vol. 19, no. 1, pp. 122–130, Jan. 2022, doi: 10.1016/j.jacr.2021.08.026.

[7] V. V. Do, E. R. Núñez, F. G. Wilder, and N. Nguyen, "Negative social determinants of health are linked to lung cancer screening underutilization," *Lung Cancer Manag.*, vol. 14, no. 1, p. 2583639, Dec. 2025, doi: 10.1080/17581966.2025.2583639.

[8] CDC, "Social Determinants of Health," Public Health Professionals Gateway. Accessed: Dec. 07, 2025. [Online]. Available: https://www.cdc.gov/public-health-gateway/php/about/social-determinants-of-health.html

[9] NCDHHS SCHS, "NCDHHS: DPH: NC SCHS: Statistics and Reports: Cancer: Incidence Rates," SCHS: Cancer Incidence Rates. Accessed: Dec. 08, 2025. [Online]. Available: https://schs.dph.ncdhhs.gov/data/cancer/incidence_rates.htm

[10] NCDHHS SCHS, "NCDHHS: DPH: NC SCHS: Statistics and Reports: Cancer: Mortality Rates," SCHS: Cancer Mortality Rates. Accessed: Dec. 08, 2025. [Online]. Available: https://schs.dph.ncdhhs.gov/data/cancer/mortality_rates.htm

[11] NCIOM, "NC County Health Data," NC COUNTY HEALTH DATA. Accessed: Dec. 08, 2025. [Online]. Available: https://nciom.org/nc-health-data-archive/map/

[12] National Cancer Institute Genomic Data Commons, "GDC Data Portal Homepage," Genomic Data Commons Data Portal. Accessed: Dec. 08, 2025. [Online]. Available: https://portal.gdc.cancer.gov/

[13] NIH, "All of Us Research Program Protocol," All of Us Research Program Protocol. Accessed: Dec. 08, 2025. [Online]. Available: https://allofus.nih.gov/article/all-us-research-program-protocol

[14] SEER, "Accessing the Data - SEER Datasets," How to Request Access to SEER Data. Accessed: Dec. 08, 2025. [Online]. Available: https://seer.cancer.gov/data/access.html

[15] D. Cutler and A. Lleras-Muney, "Education and Health: Evaluating Theories and Evidence," National Bureau of Economic Research, Cambridge, MA, w12352, July 2006. doi: 10.3386/w12352.