

Poverty Level Relationship with Cancer Mortality in U.S. Counties

Minh Doan, Elle Luo, Vamsi Katari, Jeff Houser

2025-12-09

Introduction

The relationship of socioeconomic status to accessibility in healthcare has arguably been a controversial topic within the United States¹. In particular, the disparity in accessing healthcare has largely affected millions of Americans, leading to unequal quality of care, waiting times, unresolved health issues, and survival rates². Understanding the link between poverty and cancer is important as it can help uncover patterns of limited access to healthcare, environmental exposure, and an absence of preventive services that disproportionately affect low-income populations. In addition, poverty levels also have some association with the ability to access healthcare insofar as insurance options are concerned and disproportionately affects certain racial groups within the United States. It is important to reveal these potential disparities in health outcomes among marginalized communities.

Therefore, the research questions guiding this study are:

- *What is the descriptive relationship between county-level poverty rates and cancer mortality?*
- *How does the strength of this relationship change when conditioning on health insurance coverage and racial composition?*
- *What is the relationship between the racial and ethnic composition of the counties and the cancer mortality rate? Is there a specific racial or ethnic group that has a higher or lower cancer mortality rate?*

The goal of our study is to help public health officials and policymakers better understand how poverty and inequality in healthcare contribute to disparity in cancer outcomes. By providing insight into the magnitude of this disparity, our findings can further guide more resource allocation, support early detection initiatives, and identify counties most likely to benefit from equitable health policy intervention and strategies.

Data Source & Provenance

To ensure high quality and reliability, we gathered data from two primary government sources: the Surveillance, Epidemiology, and End Results (SEER) program and the U.S. Census Bureau.

Cancer Mortality (Dependent Variable): Cancer mortality data were sourced from the National Center for Health Statistics (NCHS) and accessed through SEER*Stat. NCHS compiles death certificate data from state vital statistics offices, with causes of death coded under ICD standards. These records undergo national cross-checks to reconcile discrepancies, remove duplicates, and validate cause-of-death classifications, followed by automated consistency edits to ensure accuracy and comparability across states. Counties with fewer than ten cancer deaths are suppressed to maintain confidentiality and statistical stability. For all available counties, we used age-adjusted mortality rates, which standardize for differences in age structure so that observed variation reflects underlying health disparities rather than demographic composition. The rate is calculated using the formula below:

$$aarate_{x-y} = \sum_{i=x}^y \left[\left(\frac{count_i}{pop_i} \right) \times 100,000 \times \left(\frac{stdpop_i}{\sum_{j=x}^y stdpop_j} \right) \right]$$

Socioeconomic and Demographic Data (Independent Variables): All independent variables were sourced from the American Community Survey (ACS) 5-Year Estimates (2018–2022). The ACS collects data continuously using mailed questionnaires, phone interviews, and in-person enumeration, then aggregates five years of responses to produce stable county-level estimates. The Census Bureau conducts multi-stage quality checks, including nonresponse adjustment, imputation for missing items, sampling error estimation, and comparison against administrative benchmarks from agencies such as the Social Security Administration and IRS, where applicable. We selected the 5-year aggregate to ensure statistically stable estimates for small, rural counties that often have high margins of error in single-year datasets.

¹Kim, Y., Vazquez, C. & Cubbin, C. Socioeconomic disparities in health outcomes in the United States in the late 2010s: results from four national population-based studies. *Arch Public Health* 81, 15 (2023)

²Ohlson, Madeline (2020) "Effects of Socioeconomic Status and Race on Access to Healthcare in the United States," Perspectives: Vol. 12, Article 2.

Variables

Y-Concept: Cancer Mortality: The dependent variable is the Age-Adjusted Cancer Mortality Rate, operationalized as the number of cancer deaths per 100,000 population, standardized to the 2000 U.S. population.

X-Concept: Poverty: The primary independent variable is the Poverty Rate, operationalized as the percentage of the population living below the federal poverty line. This is calculated by dividing the count of individuals below the poverty threshold by the total population for whom poverty status is determined (sourced from ACS Table S2701).

Other Key Variables

Insurance Composition: Percent uninsured, percent covered by Medicaid or public assistance, and percent covered by private insurance.

Demographic Controls: To isolate the effect of poverty, we controlled for racial and ethnic composition using Non-Hispanic categories to prevent double-counting. Variables included the percentage of the population identifying as White (Non-Hispanic), Black (Non-Hispanic), Asian (Non-Hispanic), Native American (Non-Hispanic), and Hispanic (of any race). Note: Raw age variables were excluded from regression models as the dependent variable is already age-adjusted.

Result and Modeling

Table 1: Cancer Mortality Rate Relationship with Poverty, Health Insurance, and Race

	<i>Dependent variable:</i>		
	Cancer Mortality Rate		
	(1)	(2)	(3)
Percentage of People in Poverty	2.165*** (0.096)	1.196*** (0.150)	1.158*** (0.153)
Percentage of People on Medicaid		0.525*** (0.116)	0.293** (0.104)
Percentage of People with Private Insurance		-0.330*** (0.092)	-0.610*** (0.092)
Percentage of White Population			2.781* (1.146)
Percentage of Black Population			2.754* (1.150)
Percentage of Hispanic Population			2.264* (1.153)
Percentage of Asian Population			1.315 (1.223)
Percentage of Native Population			2.459* (1.140)
Percentage of Two or More Races			4.410*** (1.308)
Constant	132.243*** (1.366)	156.064*** (8.372)	-94.555 (115.502)
Observations	3,070	3,070	3,070
R ²	0.208	0.233	0.342
Adjusted R ²	0.208	0.232	0.340
Residual Std. Error	25.138 (df = 3068)	24.743 (df = 3066)	22.941 (df = 3060)
F Statistic	804.764*** (df = 1; 3068)	310.494*** (df = 3; 3066)	176.669*** (df = 9; 3060)

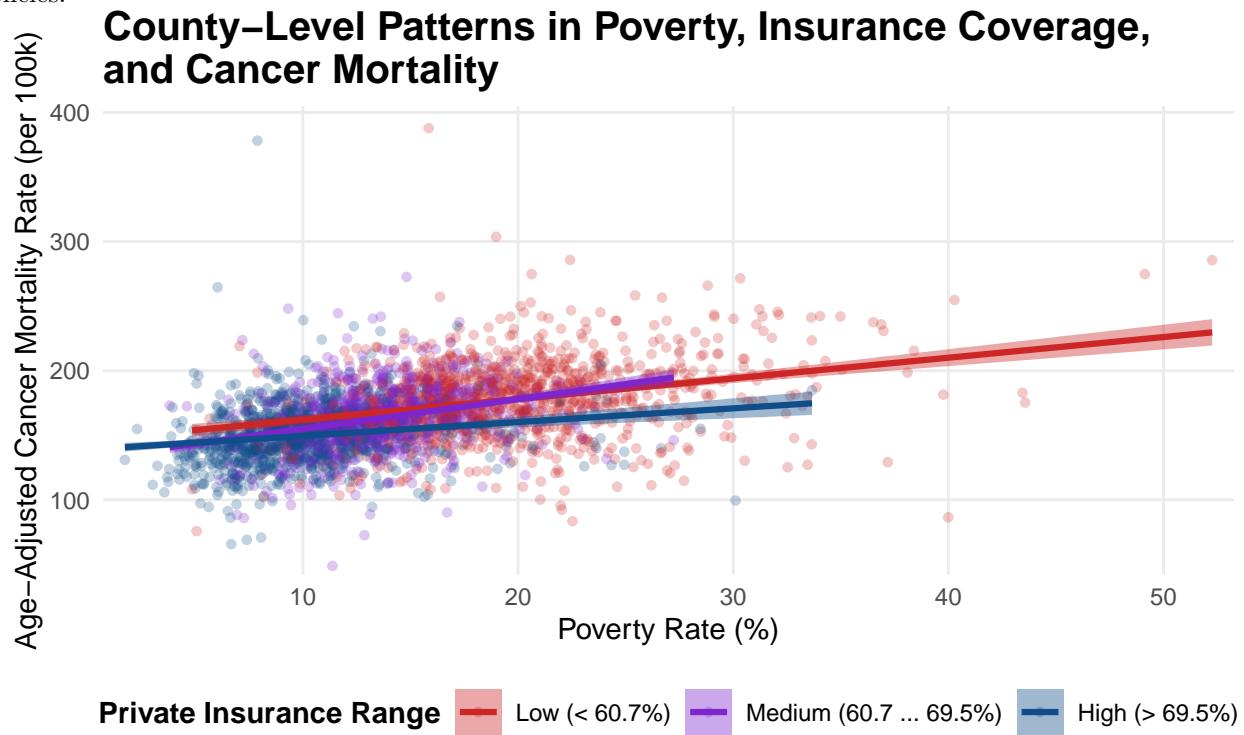
Note:

*p<0.05; **p<0.01; ***p<0.001

Model 1 - Poverty and Mortality: The simplest form of the model, comparing just the percentage of people in poverty with the cancer mortality rate, demonstrates a significant relationship between poverty and cancer mortality rate. Both the coefficient and intercept values are significant with p values below 0.001, meaning we can reject the null hypothesis that the independent variable does not have a relationship with the dependent variable and the null hypothesis that the intercept is not different from a value of 0. From a practical perspective, this means that for each ten percent increase in the number of people living in poverty within a county, the age-adjusted cancer mortality rate increases by approximately 21 deaths per 100,000 people.

Model 2 - Role of Insurance: As we expand our model to incorporate the type of insurance, we see that the coefficient for the percentage of people in poverty roughly halves but still remains significant. While the coefficients relating to insurance type also have significance, suggesting they have a relationship with cancer mortality when county poverty levels are accounted for. Practically speaking, the insurance status coefficients suggest that if the percentage of people on Medicaid within the county were to be ten percent higher, then roughly six more people per 100,000 would die due to cancer compared to a county that experiences similar levels of poverty. Conversely, a county with ten percent more people on private insurance would expect roughly three fewer cancer deaths per 100,000 people.

Model 3: Demographic factors: In the context of insurance and poverty, the ethnic composition of the county appears to have some relationship with cancer mortality outcomes. After controlling for economic factors, racial composition remains a significant predictor. It appears that counties with higher populations of people comprising two or more races have worse outcomes insofar as cancer mortality is concerned, with statistical significance. While it may be difficult to discern precise practical meaning from the racial coefficients, we thought it was still important to include this model since it has implications for what communities may be able to benefit the most from policies hoping to provide more equity in healthcare to ameliorate cancer outcomes. This finding highlights specific communities that would benefit most from targeted health equity policies.



Model Assumptions and Diagnostics

Given our sample size of approximately 3,000 U.S. counties, we rely on large-sample statistical theory to evaluate our model. This approach requires us to assess two primary conditions: (1) that the data are Independent and Identically Distributed (IID), and (2) that a Best Linear Predictor (BLP) exists and is

unique.

IID: This assumption requires that each data point is independent of the others and drawn from the same underlying distribution.

- Independence: Since our data represents counties, cancer mortality rates in one county are likely related to rates in neighboring counties due to shared environmental factors, regional diets, or state-level healthcare policies, suggesting a potential for violation of independence due to this geographical clustering.
- Identically Distributed: The assumption that all counties come from an “identical” distribution is also likely violated. Counties vary drastically in population size (e.g., Los Angeles County vs. a rural county in North Dakota). This variation suggests heteroskedasticity, meaning the variability of cancer rates is likely higher in smaller counties than in larger ones.
- The consequences: While these violations do not prevent us from estimating the relationship (our coefficients remain consistent), they do affect how we calculate uncertainty. Specifically, clustering and heteroskedasticity likely cause our standard errors to be underestimated, meaning our results may appear more precise than they actually are.

Existence and Uniqueness of the Estimator: This assumption ensures that a linear relationship can theoretically be calculated from the data

- Existence: To establish that the best linear predictor (BLP) exists requires finite covariance between our different X variables with one another and also with our Y variable. All of the X variables are percentages and are thus finite. They also do not exhibit heavy tails, suggesting finite covariance between the different variables. Our Y variable of cancer mortality rate also does not exhibit a heavy tail and conceptually represents a finite value and would have finite covariance with the different X variables. These taken together suggest the BLP exists.
- Uniqueness: To evaluate that this BLP is unique requires that there not be perfect collinearity between the different X variables. While our racial composition variables are naturally related (as the percentage of one group rises, others must fall), they are distinct enough to allow us to isolate the specific relationship of each variable. No variables were automatically dropped by our software, confirming that the model is mathematically unique.
- Assessing the Shape: To confirm that a linear model effectively describes the actual pattern in the data, we examined the Residuals vs. Fitted plot (Appendix A3). The plot shows the trend line (blue) tracking the zero line closely, without a distinct “U” shape or curve. This suggests that a straight-line relationship is an appropriate description of the data

Conclusion

This study examined how poverty, insurance coverage, and racial or ethnic composition are related to county level cancer mortality in the U.S. Across all models, socioeconomic disadvantage consistently predicts higher age adjusted mortality. A 10 point increase in poverty is associated with roughly 21 additional deaths per 100,000 people. Adding insurance variables weakens but does not remove the link between poverty and mortality, and differences in Medicaid, private insurance, and uninsured rates meaningfully shape outcomes. In the full model, racial composition, including the size of multiracial populations, remains associated with higher mortality even after accounting for poverty and insurance.

These patterns highlight persistent structural inequalities: counties with high poverty, limited private insurance access, and larger marginalized populations face disproportionate cancer burdens. The findings support targeted public health investments in prevention, early detection, treatment access, Medicaid expansion, equitable reimbursement, and improved resources for rural and minority serving facilities.

Several limitations should be noted: observational data prevent causal inference; neighboring counties may share unmeasured characteristics; ACS and SEER data may mask localized disparities; and broad racial categories cannot fully capture historical and social inequities. Future work could use multilevel, spatial, or longitudinal approaches to address these issues. Overall, this study clarifies how poverty, insurance coverage, and demographic composition jointly shape cancer mortality and offers evidence to guide more equitable health policy interventions.

Appendix

A1. Link to Data Source

Github link to datasets:

https://github.com/jhouser220/DataSci203_Lab2/tree/main/Code%20and%20Data

Census data derived from:

<https://data.census.gov>

SEER data derived using custom software from:

<https://seer.cancer.gov/>

A2. List of Model Specifications Tried (A brief list with 1-sentence takeaways, as required)

1. Extreme cases-county level: Variables: poverty rate, Medicaid share, private insurance share, and cancer mortality. Takeaway: Even among the poorest and wealthiest counties, the direction of insurance-mortality relationships remains consistent with the main model.
2. Extreme cases-state level: Variables: state-level poverty rate, Medicaid share, private insurance share, and cancer mortality. Takeaway: The insurance-mortality relationship persists when the analysis is aggregated to states, indicating the pattern is not driven solely by county-level variation.
3. Mortality rate and race: Variables: racial composition shares, cancer mortality. Takeaway: Higher Black and Native American population shares correlate with higher mortality, while higher Hispanic and Asian shares correlate with lower mortality
4. Mortality rate, Urban and Poverty: Variables: urban classification, poverty rate, cancer mortality. Takeaway: Urban counties show modestly lower mortality, and this advantage remains after controlling for poverty.
5. Morality rate, age: Variables: Pct_over_65, cancer mortality. Takeaway: Age has a known association with worse outcomes in cancer. We wanted to confirm that the age-correction of the data would show an absence of a relationship and it did.

A3. Residuals-vs-Fitted-Values Plot

Cancer Mortality Model Residuals vs Predictions

