The Relationship Between Poverty and Cancer Mortality in US Counties

MIDS W203 - Lab 2

Lizzie Friend, Andre Gigena, Chris McClure-St. Amant

7/24/2022

Contents

Introduction

Despite significant advancements in science and medicine, the Centers for Disease Control and Prevention listed cancer as the second highest leading cause of death in the United States in 2020[1]. Many, including medical groups and governments, have a vested interest in lowering mortality from cancer. These groups have made efforts to reduce risk factors for cancer, such as sponsoring advertising campaigns encouraging early screenings or emphasizing the negative effects of smoking. These efforts are commendable, but would an effort to lower the poverty rate pay dividends in lowering cancer mortality?

Poverty can influence a number of factors which may correlate with increased risk of mortality from cancer, such as stress, poor diet and nutrition, coping behaviors like smoking and substance use, and living in under-resourced areas which might involve exposure to environmental hazards. Poverty can also influence one's ability to obtain regular preventative care and screenings and, should a cancer diagnosis be given, one's ability to travel to hospitals and afford life-saving interventions.

Using a conglomeration of data sets detailing county-level factors such as poverty rates, superfund sites, air pollution levels, behavioral factors (such as smoking), and more, this analysis intends to answer the research question: What is the relationship between poverty and cancer mortality in US counties? This will be achieved using a nested set of linear regression models.

Data & Methodology

Data was sourced from clinicaltrials.gov[2], cancer.gov[2], census.gov[3,4], the US Energy Information Administration (EIA)[5], the Environmental Protection Agency (EPA)[6], the Centers for Disease Control (CDC)[7] and countyhealthrankings.org[8]. We began with a county-level data.world data set on cancer trials conducted from 2010-2016, which included 12 variables and 3,072 observations (representing most of the 3,143 United States county and county equivalents)[2]. The following variables were added via joins from additional sources, with attempts made to match the cancer trial time frame as closely as possible:

- Number of superfund sites by county (EPA)
- Average daily environmental toxicology measures by county (CDC)
- Population by county (Census.gov)
- Percent of the population that smokes by county (countyhealthrankings.org)
- Coal usage by state (EIA)

Most data fields were used in their original forms. An exception to this is the population by county variable, which was divided by 100,000 in order to reduce the decimal places of its linear regression coefficient. Additional data wrangling steps included aggregating superfund site data to create counts per 100,000 residents by county, modeling state-level coal-use data to calculate usage in short tons per 100,000 residents, and aggregating daily, county-level pollution readings to create annual averages.

[1]Centers for Disease Control and Prevention, "Leading Causes of Death," *cdc.gov*, Jan. 13, 2022. [Online]. Available: https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm. [Accessed: Jul. 28, 2022].

[2]N. Rippner, "Cancer Trials" *data.world*, Jul. 13, 2016. [Online]. Available: https://data.world/nrippner/cancer-trials. [Accessed: Jul. 14, 2022].

[3]U.S. Census Bureau, "State Population by Characteristics: 2010-2019," *census.gov*. [Online]. Available: https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-detail.html. [Accessed: Jul. 20, 2022].

[4]U.S. Census Bureau, "state.txt," *census.gov*. [Online]. Available: https://www2.census.gov/geo/docs/reference/state.txt. [Accessed: Jul. 20, 2022].

[5]U.S. Energy Information Administration, "Coal Data Browser," *eia.gov*. [Online]. Available: https://www.eia.gov/coal/data/browser/#/topic/20?agg=0,2,1&geo=vvvvvvvvvvvvo&freq=A&start=2001&end=2020&ctype=map&ltype=pin&rtype=s&maptype=0&rse=0&pin=. [Accessed: Jul. 20, 2022].

[6]U.S. Environmental Protection Agency, "LIST-008R Active Site Status Report," *epa.gov*. [Online]. Available: https://semspub.epa.gov/work/HQ/401146.pdf. [Accessed: Jul. 20, 2022].

[7]Centers for Disease Control and Prevention, "Daily County-Level PM2.5 Concentrations, 2001-2014," *cdc.gov*. [Online]. Available: https://data.cdc.gov/Environmental-Health-Toxicology/Daily-County-Level-PM2-5-Concentrations-2001-2014/qjju-smys. [Accessed: Jul. 20, 2022].

[8]County Health Rankings and Roadmaps, "National Data & Documentation: 2010-2020," *countyhealthrankings.org*. [Online]. Available: https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation/national-data-documentation-2010-2019. [Accessed: Jul. 20, 2022].

A constraining factor in our data was missing values for over 400 counties in the countyhealthrankings.org smoking data. The excluded counties tended to be those with very low populations, and they were dropped from the data set. The data was then split into two sets: 30% was used as an exploratory set (797 observations), while the other 70% was reserved for this research (1,872 observations).

Cancer mortality was defined as the number of cancer-related deaths per 100,000 residents, while poverty was operationalized by the estimated percentage of residents in each county living below the poverty level.

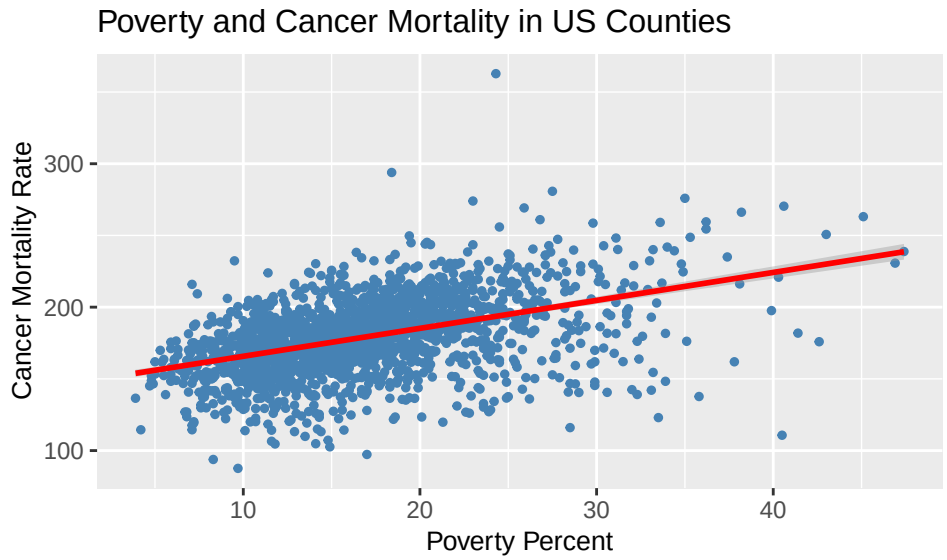

Figure 1: Poverty vs. death rate with regression line

To assess the relationship between poverty and cancer mortality, we fit a linear regression of the following form, where $\mathbf{Z}\gamma$ represents additional covariates included to increase precision:

$$\widehat{cancer\ mortality} = \beta_0 + \beta_1 \cdot poverty + \mathbf{Z}\gamma$$

These additional covariates included:

- **Cancer incidence**: Defined as cancer cases per 100,000 residents, included to control for the effect that higher incidence could have on higher mortality rates.
- **County population**: Included as a proxy to control for access to larger/more equipped hospitals; our assumption is more populated counties will have better medical facilities.
- **Air quality**: Defined by daily county-level readings of particulate matter concentrations, averaged across 2015 and included to control for air pollution that may increase risk of cancer.
- **Superfund sites**: Defined as superfund sites per county, per 100,000 residents, and included as a proxy for environmental pollution that is not airborne.
- **Tobacco use**: Defined by the percentage of the county population that smokes and included to control for behavioral factors that increase risk of cancer.
- **Coal use**: Defined as coal usage by state per 100,000 people and included as an additional control for pollution.
- **State**: One-hot encoding of state, which controls for all other omitted state-level factors (e.g. increased risk of skin cancer due to climate).

There are other factors that could be beneficial to control for, such as genetic predisposition and access to healthy foods; however, we were unable to find good representations of these data in the allotted time.

Results

Table 1 shows the results of our five regression models. In all five, the coefficient on poverty rate was highly statistically significant ($p < 0.001$). As covariates were added to control for pollution, location, population, and other factors, the poverty coefficient decreased from 1.95 to 1.06. F testing of all models suggested improved precision with each level of additional covariates, indicating model five performs best. Model five suggests a 1% increase in the percentage of county residents living below the poverty line is predicted to increase cancer deaths by approximately 1 per 100,000 residents. The coefficients on both air quality and tobacco use are also highly significant.

The average cancer mortality rate across all counties in the test data was 179 per 100,000 residents, so the practical implication of this result is a 2% reduction in poverty rate is predicted to reduce cancer deaths by 1%. Estimates suggest between 500,000 and 600,000 cancer deaths occurred in the US in 2015[9], so this model predicts a 2% poverty reduction could have saved 6,000 lives in 2015 alone.

Table 1: Estimated Regressions

	Output Variable: cancer deaths per 100k people				
	(1)	(2)	(3)	(4)	(5)
Poverty	1.95***	1.87***	1.16***	1.12***	1.06***
	(0.12)	(0.09)	(0.09)	(0.09)	(0.10)
Cancer incidence		0.25***	0.21***	0.19***	0.20***
		(0.01)	(0.01)	(0.01)	(0.01)
County population		−0.73**	−0.30*	−0.51**	−0.37*
		(0.22)	(0.12)	(0.15)	(0.14)
Air quality				2.55***	1.26**
				(0.25)	(0.48)
Superfund sites				0.04	0.10
				(0.05)	(0.05)
Tobacco use			219.70***	196.20***	170.40***
			(10.63)	(11.15)	(11.91)
Constant	146.30***	37.11***	24.73***	15.49***	23.73*
	(1.84)	(4.92)	(4.09)	(4.32)	(9.59)
coal use				✓	✓
state					✓
Observations	1,872	1,872	1,872	1,872	1,872
R^2	0.20	0.43	0.55	0.58	0.62
Residual Std. Error	24.00 (df = 1870)	20.24 (df = 1868)	18.04 (df = 1867)	17.52 (df = 1864)	16.75 (df = 1818)
Note:	*p<0.05; **p<0.01; ***p<0.001				
	HC_1 robust standard errors in parentheses.				
	Additional features are state and coal use by state				

[9]R. Siegel, K. Miller, and A. Jemal, "Cancer Statistics, 2015," *CA, A Cancer Journal for Clinicians*, vol. 65, no. 1, pp. 5-29, January 2015. [Online Serial]. Available: https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21254. [Accessed: Jul. 29, 2022].

Limitations

To evaluate the consistency of our regression estimates, we evaluate the independent and identically distributed (iid) assumption. The scope of our data set–including most counties–suggests it does meet the iid requirement, with a few caveats. The fact that select very small counties were missing data and thus excluded suggests the set favors larger counties, which may differ from smaller counties in ways that could affect our results. Similarly, the approximately 60 counties omitted included in the original data set may represent an unknown cluster (e.g. lower funding or unique governance factors resulting in the data not being collected or reported). Further analysis would be required to assess common factors among the missing counties and determine how their omission may have affected our model.

Another assumption we evaluate in regards to the consistency of our regression estimates is that a unique best linear predictor exists and describes the population distribution, meaning no perfect collinearity exists between variables. Figure 2 shows a correlation heatmap of input variables, finding no perfect or high collinearity. All variables were also checked visually for heavy tails, and while population and superfund sites were right-skewed, we do not believe them to be heavy-tailed.

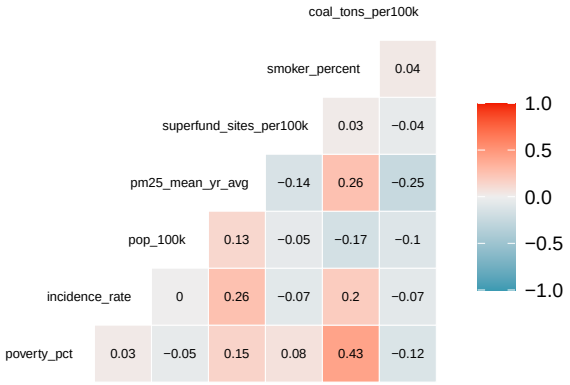

Figure 2: Correlation heatmap

Causes of cancer are both numerous and not fully understood by medical science[10]. As such, there are many potential omitted variables that could bias our estimates. Our model includes controls for several environmental factors, but we were unable to control for genetics, and tobacco use was the only included lifestyle factor. Some cancers are more deadly than others, so a genetic predisposition to certain cancers could drive up death rates independent of other factors, likely causing the coefficient on poverty to be higher than its true value and biasing the estimate away from zero. Analogous reasoning applies to other potential omitted variables, including lifestyle- and medical history-related variables.

The data used in this study came from numerous sources and it was not always possible to precisely match timeframes. There is no time series data used in this analysis, and to the best of our knowledge, all data sources for output and inputs to our model are from the years 2010-2016 (with the exception of the superfund data, which came from 2022). Because the measures used are not likely to change drastically over the time period evaluated, we do not see this as a significant limitation.

Conclusions

This analysis focused on the relationship between county-level poverty and cancer deaths. For an increase of 1% in poverty rate, our models estimate an increase in cancer death rate between 1.06 and 1.95 people per 100k. We note that the coefficients on air quality and tobacco use were highly significant as well, but due to potential omitted variable bias, we do not propose a practical interpretation of their meaning. Future research could benefit from a smaller unit of observation (perhaps individuals), to allow for more precise controls for genetic and lifestyle factors. The ultimate goal of this research is to increase our understanding of the specific negative effects of poverty on cancer patients, and lend important detail to the already rich body of evidence on the broader effects of poverty on American people and communities.

[10]Stanford Healthcare, "What Causes Cancer?," *stanfordhealthcare.org*. [Online]. Available: https://stanfordhealthcare.org/medical-conditions/cancer/cancer/cancer-causes.html. [Accessed: Jul. 29, 2022].