The Role of Location in Health Outcomes: A Comparison of Urban and Rural California Census Tracts

Thanh Mai
University of California, Los Angeles

I. INTRODUCTION

Urban-rural disparities in health outcomes have been an important area of research due to the diverse factors that shape health in different settings[1]. For instance, rural areas often face worse health outcomes[2] due to limited healthcare access and higher poverty rates. However, urban areas also encounter challenges such as air pollution[3] or overcrowding, which can negatively impact health as well. These urban-rural health dynamics are complex, with varying implications for public health policy and interventions. Understanding these disparities are crucial to addressing location-based health inequities.

This report explores the differences in various health outcomes prevalence rates between urban and rural census tracts in California but focuses primarily on cancer and heart disease. These two diseases were selected for further analysis because they are currently two of the leading causes of death in the United States[4] and understanding their distribution across urban and rural areas is important for public health interventions. This analysis aims to answer two questions: 1) whether there is a significant difference in various health outcome prevalence rates between urban and rural California census tracts, and 2), whether urban-rural classification continues to significantly impact prevalence rates after controlling for confounding variables such as sex, age, poverty, education, and health insurance coverage.

II. METHODS

A. *Data Sources*

Health outcomes prevalence data was obtained from the Centers for Disease Control and Prevention (CDC) PLACES Data Portal[5]. The CDC PLACES dataset includes prevalence rates on thirteen different health outcomes, ranging from arthritis to diabetes at the census tract level. Demographic and socioeconomic counts for variables including sex, age, poverty, education, and health insurance were obtained from the United States Census Bureau, through their Decennial Census Survey[6] and 5-Year American Community Survey[7]. These counts were aggregated to the census tract level in order to calculate the following: the proportion of males and females (sex), the proportion of adults aged 65 and over (age), the proportion of individuals with a bachelor's degree (education), the proportion of the population below 200% of the federal poverty level (poverty), and the proportion of individuals with health insurance coverage (insurance). Since the data from the CDC is based on the adult population (18+), sex and age groups calculated in this study are also based on the adult population in each census tract for consistency. Proportions (rates) were used instead of categorical variables (i.e., male vs. female, insured vs. uninsured, etc.) to ensure that the analysis accounts for differences in population size across census tracts and retains greater detail in the data.

B. *Study Population*

The study population consists of adults aged 18 and older residing in the 9,129 census tracts across California.

C. *Statistical Methods*

Descriptive statistics were initially examined to provide an overview of health outcomes across urban and rural census tracts. A Welch's two-sample t-test was subsequently conducted to assess whether statistically significant differences existed in the prevalence rates of various health outcomes between urban and rural areas, with statistical significance defined at an alpha level of 0.05. To further investigate these disparities, cancer and heart disease were selected for more detailed analysis. Multiple linear regression analysis was utilized to evaluate the influence of urban-rural

Table 1. Sociodemographic Summary Statistics

	Rural (%)	Urban (%)
Female	53.047	48.695
Male	46.953	51.305
Age 65+ (Older Adults)	26.422	19.189
Age 18-64	73.578	80.811
Insured	89.793	89.761
Below Poverty	31.759	29.286
Bachelor's Degree	15.517	21.972

classification on cancer and heart disease, separately, while adjusting for potential confounders, including sex, age, poverty, education, and health insurance status. The regression models were evaluated through a series of diagnostic checks, including an assessment of model fit via adjusted R-squared, multicollinearity tests, and correlations among predictors. Additionally, residual plots were examined to evaluate assumptions of normality, linearity, and homoscedasticity. All analyses were conducted in R Studio version 4.4.1.

III. RESULTS

A. *Sociodemographic Descriptive Statistics*

The sociodemographic descriptive statistics presented in Table 1 provide an overview of the mean values for key variables across urban and rural areas. Rural areas tend to have a higher proportion of older adults (26.4% vs. 19.2%) and females (53.0% vs. 48.7%). Additionally, compared to their urban counterparts, rural areas also have slightly higher poverty rates (31.8% vs. 29.3%) and lower bachelor's degree attainment (15.5% vs. 22.0%), while insurance coverage is nearly identical between the two groups (approximately 89.8%).

Differences in prevalence rates are evident in Figure 1, with rural areas generally reporting higher prevalence rates across all thirteen health outcomes. These descriptive statistics offer essential context for understanding the distinctions between urban and rural groups in terms of health outcomes and sociodemographic factors.

B. *Welch's Two Sample T-test Results*

To identify if the differences in health outcomes prevalence rates between urban and rural tracts illustrated in Figure 1 were statistically significant, a Welch's two-sample t-test was conducted. The results revealed significant differences in prevalence rates between urban and rural census tracts ($p<0.001$), with rural areas exhibiting higher average prevalence rates for all health outcomes, including both cancer and heart disease.

C. *Cancer Linear Regression Results*

Following the t-test analysis, cancer was selected for further investigation due to being a leading cause of death in the country. The multiple linear regression model found that cancer prevalence rates were significantly lower in urban areas compared to rural areas, even after accounting for sociodemographic factors such as age, sex, education, poverty, and insurance (Table 2). The model explained approximately 88% of the variance in cancer prevalence (R^2=0.875), indicating a good model fit. In addition to urban-rural classification being a significant predictor, older age and health insurance coverage were also associated with increased

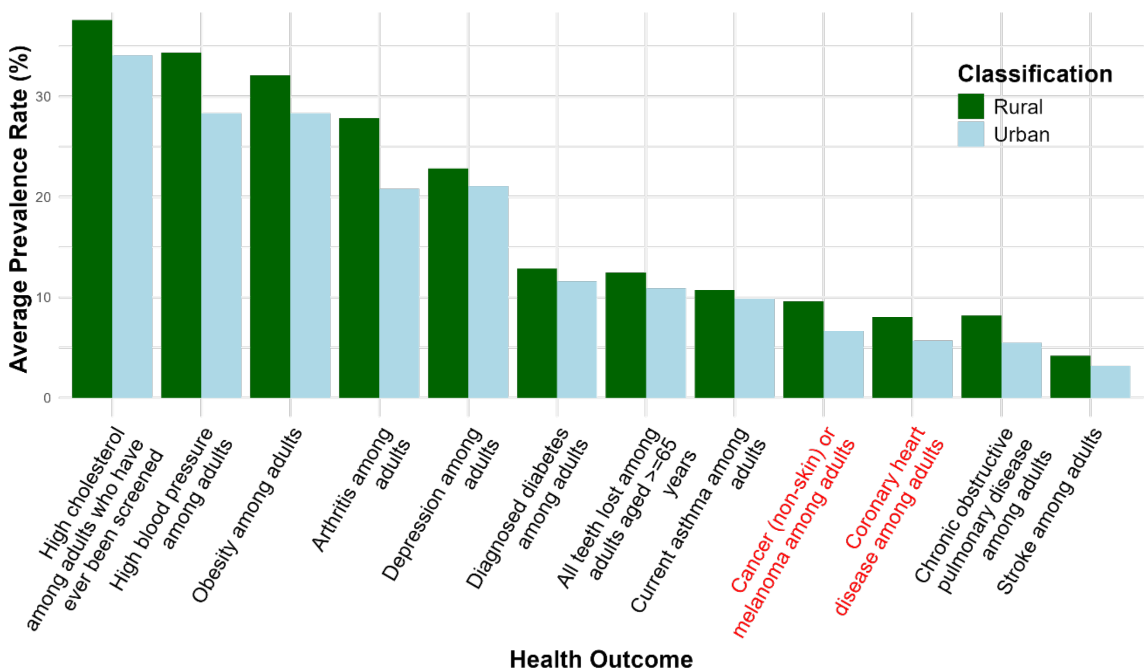

Figure 1. Histogram of Average Prevalence Rates (%) for Various Health Outcomes by Urban-Rural Classification

*Cancer and heart disease are highlighted red as these outcomes were selected for further analysis.

prevalence rates, aligning with existing literature that highlights the role of age as a risk factor and the potential for higher detection rates among insured individuals due to greater access to healthcare.

To confirm the reliability of the model, a correlation matrix (Figure 2) was created to examine the relationships among predictors, revealing no concerningly high correlations. Furthermore, multicollinearity among predictors was assessed using variance inflation factors (VIF), all of which were below 2.3, confirming that multicollinearity was not an issue in this model.

However, visual inspection of residual plots for this linear model suggested potential heteroscedasticity, which was also confirmed by a significant Breusch-Pagan test ($p<0.001$). To account for this, robust standard errors were applied, and all original significant predictors, including urban-rural classification, remained significant.

D. *Heart Disease Linear Regression Results*

Similar to cancer, the linear regression model for heart disease demonstrated that urban areas also had significantly lower heart disease prevalence compared to rural areas, even after controlling for potential confounding factors (Table 3). This model explained approximately 86% of the variance in heart disease prevalence ($R^2=0.8589$). Older age and poverty were also strongly associated with higher prevalence, while higher educational attainment was linked to lower prevalence. In contrast to the cancer linear regression model, health insurance rate was not found to be a significant predictor in the heart disease model, suggesting that other factors may better explain variations in heart disease outcomes.

The correlation matrix (Figure 2) showed no concerning correlations between the predictors and heart disease, and variance inflation factors (VIF) were all below 2.3, suggesting that multicollinearity did not present as an issue for this model.

Visual inspection of residual plots for this linear model also suggested potential heteroscedasticity, which was then confirmed by a significant Breusch-Pagan test ($p<0.001$). When applying robust standard errors to account for this, urban-rural classification and all other originally significant predictors remained significant.

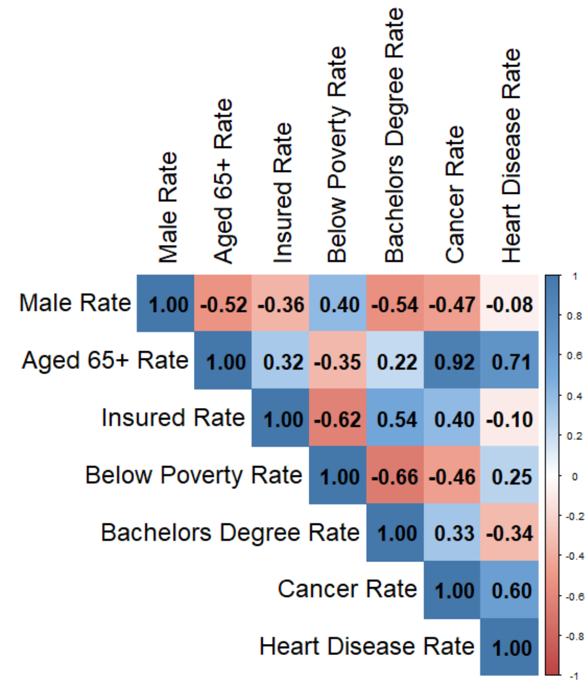

Figure 2. Correlation Matrix for All Numeric Predictors and Cancer/Heart Disease Rates

IV. CONCLUSIONS

The results demonstrated that urban and rural areas exhibit significant differences in the prevalence rates of all thirteen health outcomes, with rural areas

Table 2. Cancer Linear Regression Model

	Estimate	Std. Error	P-value
Intercept	0.0316668	0.0022100	p < 0.001
Urban Census Tract	-0.0096130	0.0004480	p < 0.001
Male Rate	-0.0259607	0.0032490	p < 0.001
Age 65+ Rate	0.2507989	0.0012999	p < 0.001
Bachelor's Degree Rate	0.0166007	0.0011630	p < 0.001
Below Poverty Rate	-0.0160766	0.0008247	p < 0.001
Health Insurance Rate	0.0115091	0.0016886	p < 0.001
Model: Cancer Rates ~ Urban Census Tract + Male Rate + Age 65 and Older Rate + Bachelor's Degree Rate + Below Poverty Rate + Health Insurance Rate			
$R^2 = 0.875$		Adjusted $R^2 = 0.8749$	

Table 3. Heart Disease Linear Regression Model

	Estimate	Std. Error	P-value
Intercept	0.05373	0.001476	$p < 0.001$
Urban Census Tract	-0.004982	0.0002993	$p < 0.001$
Male Rate	-0.05281	0.00217	$p < 0.001$
Age 65+ Rate	0.1719	0.0008682	$p < 0.001$
Bachelor's Degree Rate	-0.03866	0.0007768	$p < 0.001$
Below Poverty Rate	0.0348	0.0005508	$p < 0.001$
Health Insurance Rate	0.00008641	0.001128	$p = 0.939$
Model: Heart Disease Rates ~ Urban Census Tract + Male Rate + Age 65 and Older Rate + Bachelor's Degree Rate + Below Poverty Rate + Health Insurance Rate			
$R^2 = 0.8589$		Adjusted $R^2 = 0.8588$	

generally showing higher prevalence rates. Further investigation into cancer and heart disease revealed that urban-rural classification remains a significant predictor for both conditions, even after controlling for confounding variables such as sex, age, poverty, education, and health insurance coverage. These findings highlight the ongoing influence of urban-rural classification on health outcome prevalence rates, emphasizing its importance as a factor in public health disparities across California.

V. DISCUSSION

A. *Key Notes*

The findings of this analysis align with current research that suggests urban-rural disparities in health outcomes, particularly in chronic conditions like cancer and coronary heart disease. The consistent pattern of lower prevalence rates in urban areas underscores the complex relationship between geographic location and health outcomes. This research is important because it highlights the need for targeted public health interventions aimed at addressing these disparities, especially in rural areas, where healthcare access and resources may be more limited. By understanding how urban-rural classification impacts the prevalence of diseases like cancer and heart disease, policymakers and healthcare providers can design more effective strategies to mitigate these disparities and improve health outcomes for underserved populations.

B. *Limitations*

This analysis provides valuable insights into urban-rural health disparities, but there are several ways to improve the model and broaden its applicability. First, while the linear regression models showed good fit, further refinement could enhance the results. For example, while robust standard errors were applied to address heteroscedasticity, exploring alternative models, such as logarithmic or polynomial transformations, may help better capture non-linear relationships and improve model accuracy. Additionally, the current models only include sex, older age, poverty, bachelor’s degree attainment, and health insurance coverage. Incorporating other potential confounding variables, such as housing, mental and physical health, and broader healthcare access, could provide a more comprehensive picture of the factors influencing these disparities.

C. *Future Research*

Future research could expand the analysis beyond California to include a larger range of states. California has a higher proportion of urban census tracts compared to rural ones, which may limit the generalizability of the findings. By incorporating data from other states, researchers could obtain a more representative sample of both urban and rural populations, potentially offering a clearer understanding of urban-rural health disparities across the country.

In conclusion, this analysis illustrates that urban-rural classification of a census tract plays a significant role in prevalence rates for various health outcomes, including cancer and heart disease. These results highlight the importance of considering urban-rural status in public health interventions, particularly for rural areas where vulnerable populations face compounded risks.

REFERENCES

1. Cross, S. H., Califf, R. M., & Warraich, H. J. (2021). Rural-Urban Disparity in Mortality in the US From 1999 to 2019. JAMA, 325(22), 2312–2314. https://doi.org/10.1001/jama.2021.5334
2. Singh, G. K., & Siahpush, M. (2014). Widening Rural–Urban Disparities in Life Expectancy, U.S., 1969–2009. American Journal of Preventive Medicine, 46(2), e19–e29. https://doi.org/10.1016/j.amepre.2013.10.017
3. Strosnider, H., Kennedy, C., Monti, M., & Yip, F. (2017). Rural and Urban Differences in Air Quality, 2008-2012, and Community Drinking Water Quality, 2010-2015—United States. Morbidity and Mortality Weekly Report. Surveillance Summaries (Washington, D.C.: 2002), 66(13), 1–10. https://doi.org/10.15585/mmwr.ss6613a1
4. Centers for Disease Control and Prevention National Center for Health Statistics. FastStats. (2024). https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm
5. Centers for Disease Control and Prevention. PLACES: Local Data for Better Health. https://www.cdc.gov/places. Accessed [11/20/24].
6. U.S. Census Bureau. 2020 Decennial Census Data. (2020). https://api.census.gov/data/2020/dec/pl/variables.html. Accessed [12/13/2024].
7. U.S. Census Bureau. American Community Survey 5-Year Estimates. (2020). http://api.census.gov/data/2020/acs/acs5. Accessed [12/10/2024].

APPENDIX

A. *Source Data File*

CDC Places Data - California Census Tracts (2020)

- File name: cdc_places_ca_ct.csv
- Description: This data file was sourced directly from the CDC Places Data Portal, where it was filtered to only include California census tracts. For this analysis, the data was further filtered to focus on health outcomes only.

B. *R Data Sets*

ACS Census Data (2020)

- File name: sociodemographic.csv
- Description: This data was pulled from the ACS 5-Year Estimates using the *tidycensus* package in R. The variables pulled for this data include counts for poverty, education, and health insurance by census tract.

Decennial Census Data (2020)

- File name: urban_rural.csv
- Description: This data was pulled from the Decennial Census using the *tidycensus* package in R. Main variables pulled were the urban and rural classification variables, along with counts for sex and age by census tracts.

C. *R Code*

- File name: Biostats203A_FinalProject.Rmd
- Description: This R Markdown contains code that reads in and pulls all source data (CDC PLACES, ACS 5-Year Estimates, and Decennial Data). After cleaning up the data, merging, and outputting summary statistics, we use the data to conduct t-tests on urban-rural differences in health outcomes prevalence rates. In addition, we filter to cancer and heart disease specifically to create linear regression models that account for other potential confounding factors. Then, we run code to check for model fit and residual diagnostic plots. This R Markdown also contains code to create all figures that were included in this report.