DEMOGRAPHIC AND GEOGRAPHIC FACTORS AFFECTING PROSTATE CANCER

Applied Topology project
Warsame Mead, Alex Musselman, Scott Singer

Abstract

Federal and State organizations monitor the racial and age-adjusted rates for incidence and mortality for cancers, particularly Prostate Cancer. The objective of this project is to assess mortality and incidence rates for men diagnosed with Prostate Cancer from 2014-2018 at the county level and compare factors associated with each rate, such as socioeconomic, racial, and geographic disparities. The location of interest is in Washington State and Oregon. Using data gathered from the US Census Bureau, Oregon State Cancer Registry (OsCAR) and Washington State Cancer Registry, we conducted exploratory analysis using the statistical programming language R, ArcGIS, and Jupyter notebooks. Population shows an overall linear correlation with number of prostate cancer cases, which suggests that total county population has little effect on prostate cancer rates. Our most significant finding was geographic clusters of high and low cancer rates. Other factors such as race and poverty seem to have little to no effect on prostate cancer rates at the county level.

Background and Motivation

Prostate cancer remains the most common cancer diagnosed among men in the United States, with nearly 200,000 diagnoses per year.It is imperative to investigate factors that may be correlated with Prostate Cancer. A study shows that men with Prostate Cancer in the U.S. have higher mortality rates in "more deprived groups and rural areas (Singh,24)." Also, research shows that "Men as young as 17 years are experiencing an increasing incidence of carcinoma of the prostate in much of the world (Bleyer, 56)."

Methods

> Gathered county census data and cancer data

> Cleaned and combined data into a single table using Jupyter Notebooks, Excel, and R.

> Looked for clusters in the data using Kmapper

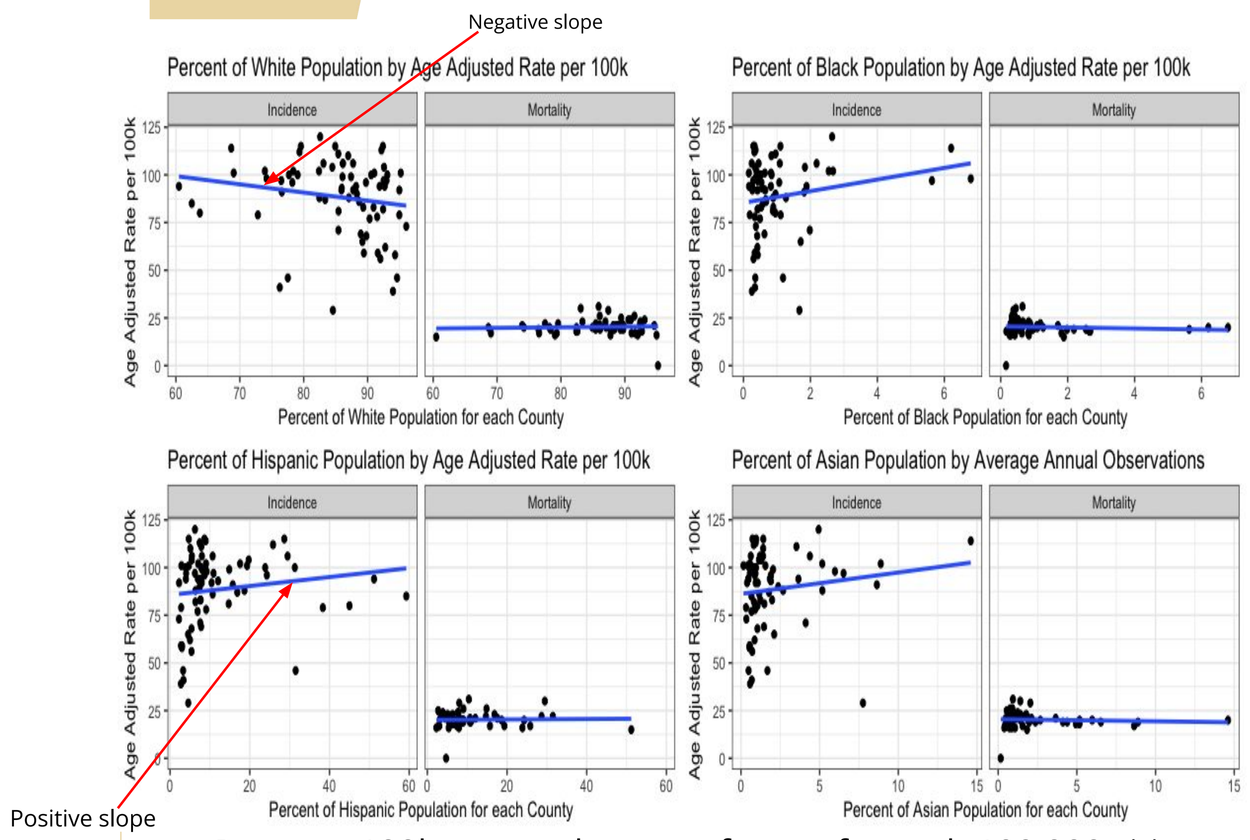
> Used Python to create scatter plots and other visualizations in Jupyter Notebooks.

> Used linear and multivariate regression techniques in R to find trends and correlations in data.

> Joined cancer and census table data to spatial data using ArcGIS Pro to analyze geographic and spatial patterns in the data.

> Used ArcGIS Pro to calculate spatial autocorrelation using Moran's I statistic.

Race



> Rate per 100k means the rate of cases for each 100,000 citizens in the population. Age adjusted accounts for different age groups compared as one.

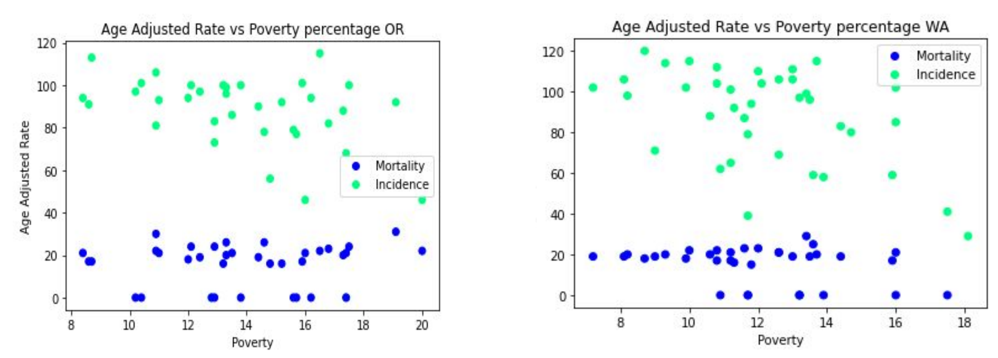
> The scatterplots are separated by Mortality and Incidences.

> Each dot represents a county within Washington and Oregon. A limitation is 19 points were removed since they had no data inputted.

> The line represents a Line of Best Fit from Linear Regression to show a linear relationship between Race and Age Adjusted Rate per 100k.

> Results: The graph depicts similar trends. Mortality rates are lower within each race group compared to incidence. Also, there appear to be a lower percent of Asian and Hispanic populations in Washington and Oregon. For the multivariate regression model, the R-squared is 0.007569.

Poverty



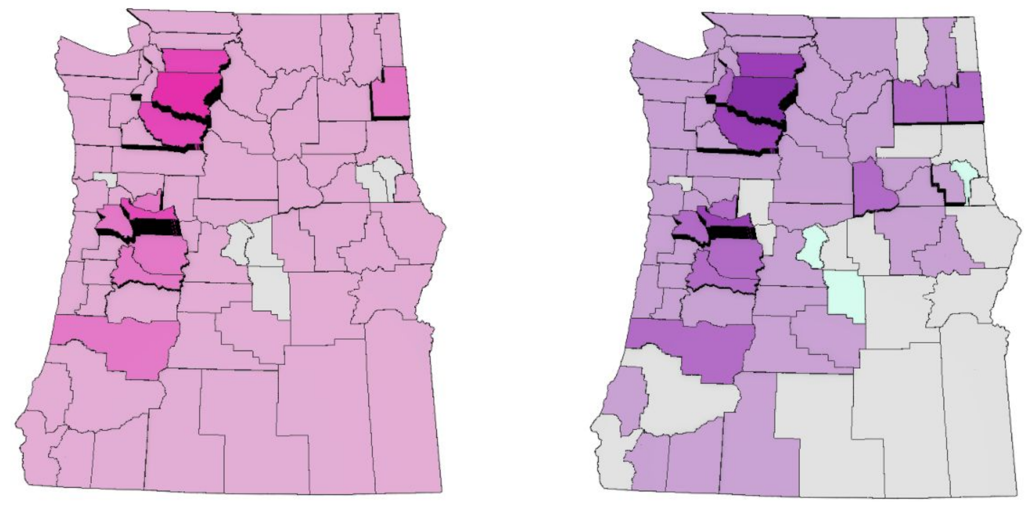
The two graphs above show the age adjusted rate of prostate cancer vs the percent of people below the poverty line for each county, blue being mortality and green being incidence. we can see a large amount of variability and a slight negative correlation in incidence, however mortality seems to be stable across poverty rates.

Limitations of Study

The data was collected by The Washington State Cancer Registry and the Oregon State Registry, because the data was collected by separate entities the data could have discrepancies due to differing methods of collection that are unexplained by our analysis.
Data was aggregate data collected at the county level, which may explain the negligible impact of race and other demographic factors in this study when compared to previous research.

Population

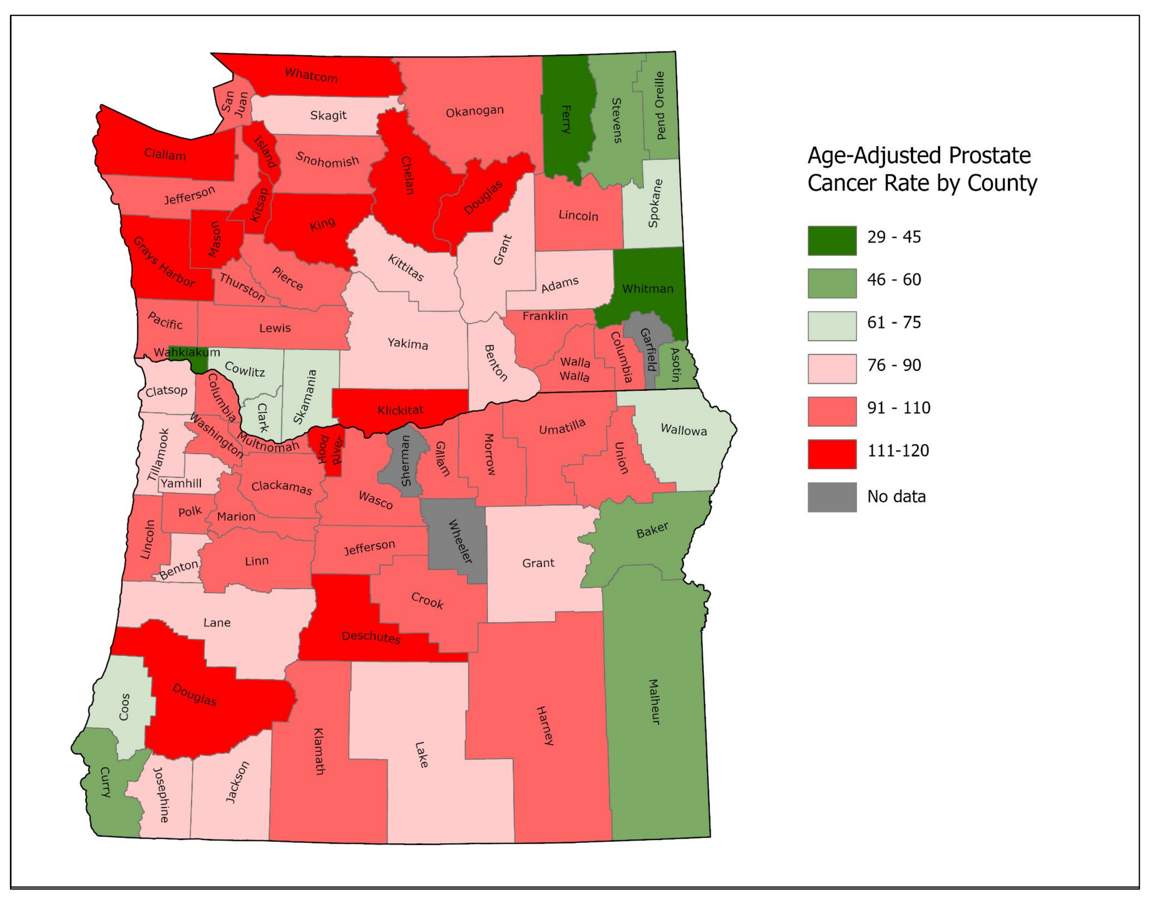
Population is one of the demographic factors we are investigating for this project. The highest concentration of population in both Washington and Oregon is located along a north-south line in the region directly west of the Cascade Mountains. When compared to the average number of cases of prostate cancer recorded by county annually, we found that the number of cases in each county showed a tendency to be proportionate to the population of the county, with a few exceptions, most notably in the eastern counties of both states.



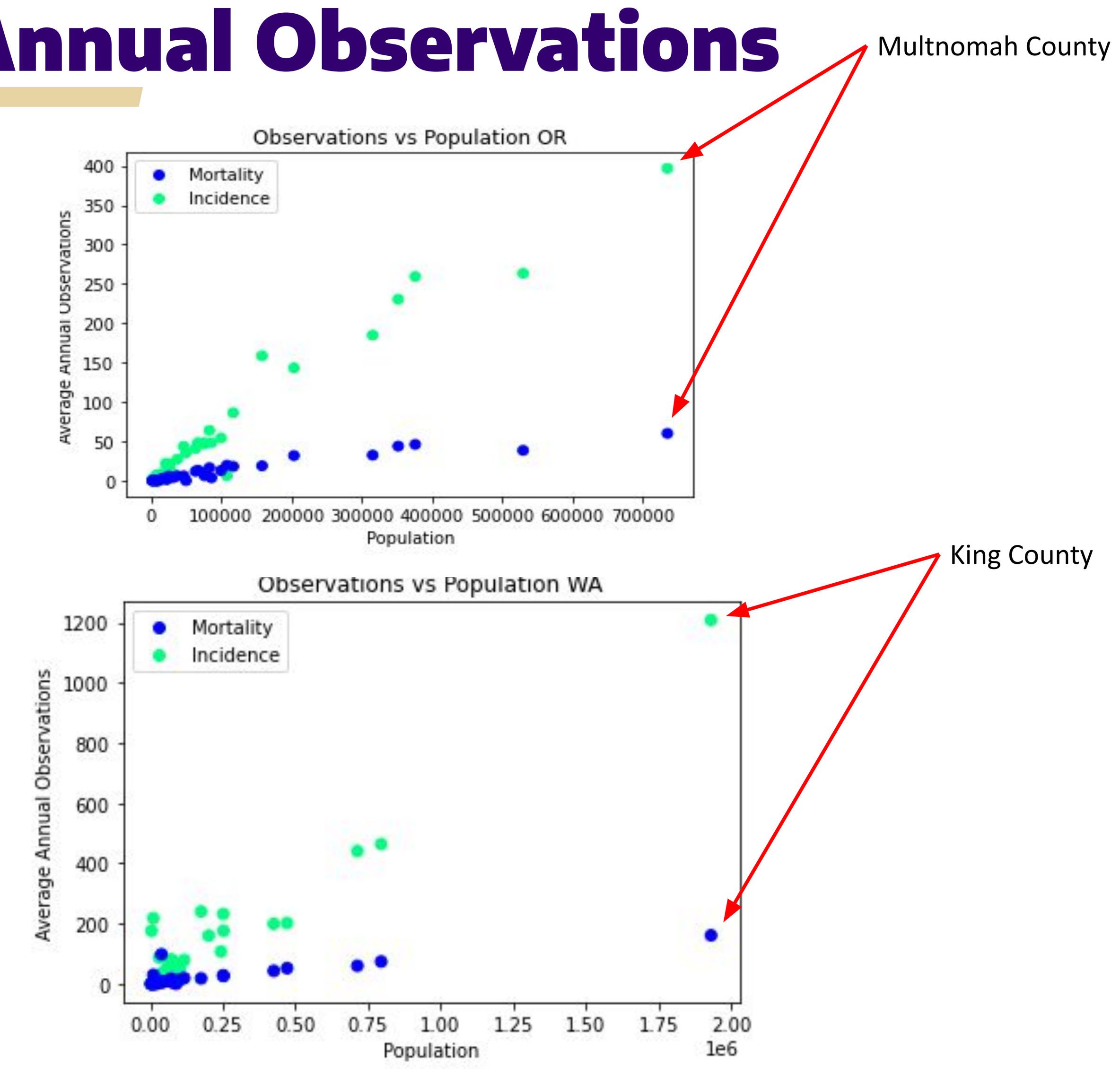
The map on the left shows population density and total population for each county in Washington and Oregon. The density of each county is represented by the height, and the total population is represented by color (gray counties have low total population, and pink counties have high total).
The map on the right shows prostate cancer incidence density and average incidence of prostate cancer per year using the same symbology, except in purple (gray counties have low average annual cases, and purple counties have high average annual cases). Light blue counties indicate missing data.

Spatial Distribution

Spatially, the highest age-adjusted rates of prostate cancer tend to occur in the western parts of Oregon and Washington, with the highest rates occurring mostly in the Puget Sound region. The eastern parts of Washington and Oregon have fairly low rates. The age-adjusted rates for Oregon and Washington are fairly similar, though Washington has more counties with very high rates. When analyzed using the Moran's I statistic with a fixed distance of 75 miles, the data was found to have a Moran's I score of .2378, indicating weak positive spatial clustering.



Annual Observations



When looking at cancer rates when compared to populations of counties we see that, the largest counties have the highest average number of cancer cases. We also see that there are about a sixth as many mortalities as there are incidences in each county regardless of population.
In the graphs above the the data shown is observations vs population. The green dots are incidence and the blue dots are mortalities. It can be seen that in higher the population the higher the amount of incidences and mortalities.
King County is an outlier in both population and cases, with almost 2 million people it has almost 3 times the amount of cases as the top county in Oregon.

Conclusion

Our hypothesis for this study was that demographic and geographic factors have a causal relationship with prostate cancer rates in Washington and Oregon. However, we have found that at the county level there is little to no correlation between these factors and prostate cancer rates, leading to a rejection of our hypothesis. We did find some interesting patterns in the spatial distribution that may warrant further analysis.

Sources

Bleyer, A., Spreafico, F., & Barr, R. (2020). Prostate cancer in young men: An emerging young adult and older adolescent challenge. *Cancer*, *126*(1), 46–57. https://doi.org/10.1002/cncr.32498
Singh, G. K., Williams, S. D., Siahpush, M., & Mulhollen, A. (2011). Socioeconomic, Rural-Urban, and Racial Inequalities in US Cancer Mortality: Part I-All Cancers and Lung Cancer and Part II-Colorectal, Prostate, Breast, and Cervical Cancers. *Journal of Cancer Epidemiology*, *2011*, 107497–27.
Oregon State Cancer Registry (1995-2023).(https://www.oregon.gov/oha/PH/DISEASESCONDITIONS/CHRONICDISEASE/DATAREPORTS/Pages/Cancer-Incidence.aspx)[Data Set] "These data were collected by the Oregon State Cancer Registry participating in the National Program of Cancer Registries (NPCR) of the Centers for Disease Control and Prevention (CDC)."
United States Census Bureau (2010-2021).
US Census community survey, 2010. US Census TIGER Lines shape files, 2020. US Census population estimates, 2020-2021
Washington State Incidence Cancer Data: Washington State Department of Health, Washington State Cancer Registry, released in April 2022 / Washington State Cancer Mortality Data: Washington State Department of Health, Washington State Cancer Registry, released in April 2022 https://doi.org/10.1155/2011/107497

Acknowledgments

Scott and Warsame would like to thank Dr. Maureen Kennedy for supporting us through our capstone project and other research we have conducted at UWT and for teaching us the skills in R and regression modeling we needed to complete this project.
Scott would also like to thank Professor Gregory Lund, Dr. Matthew Kelley, and Dr. Jim Thatcher for the training and support they provided in geographic information systems and cartography.
Further thanks goes to Monica Hust, Katie Lowe, and Mahtoska Weathers for providing useful feedback on our project early in the process.