Mapping Male Lung Cancer Risk in Washington State

Walker Azam
STAT/CSSS 554 Course Project
Date: March 13, 2023

Introduction

In this project I will be mapping and predicting Lung Cancer mortality in WA state in male populations. Lung cancer remains one of the most common forms of cancer to develop in both men and women (cancer.org, 2023). According to the Washington State Cancer Registry, Lung and Bronchus cancer had the highest mortality count, making up 21.7% of all cancer related mortalities in 2018, despite a 10.5% incidence count. It is also shown that men have higher age-specific rates of lung cancer, often due to difference in smoking habits (Jemal et al. 2003). Given this I believe performing disease mapping for Lung and Bronchus cancer in Washngton State can find regions of high risk, and can help plan public health interventions, such as encouraging early screening for Lung Cancer in men.

I will specifically be looking at cancer deaths in males, per county, aggregate counts from 2015-2019. The specific site group for cancer as defined by SEER (The National Cancer Institute's Surveillance Epidemiology and End Results) is Lung & Bronchus (Recode 22030). The data was collected from the Washington Tracking Network (more information in the **Data Background** section). The primary goal of this project is to find suitable models for disease mapping, and to provide discussion on regions of high risk, the validity of the models selected, and covariate analysis with health accessibility metrics. As aforementioned, health outcome's of smoking have been well established in regards to lung cancer, but I was particularly interested in whether health accessibility also had associations with lung cancer morbidity. In particular, this could inform whether Washington state resources could be allocated better to regions of high risk.

Data Background

Shapefiles

Washington State county shapefiles are readily available online. The shapefiles used for this study were from OpenData Washington State Department of Natural Resources (WADNR), available here: https://geo.wa.gov/datasets/wadnr::wa-county-boundaries/explore?location=47.182649%2C-120.817600%2C7.57.

A reference of all county names and locations is included in the Supplementary Materials (Supplementary Figure 1). County names will be used in this study to highlight areas of particular interest, but general locations relative to WA state will also be provided to highlight any spatial trends found.

Cancer Mortality Data

I will be primarily be working with Lung and Bronchus cancer deaths in males, per county, in WA state from 2015-2019. This data is retrieved from the Washington State Department of Health's source for environmental public health data the Washington Tracking Network (WTN). The mortality counts are collected from WA state death certificates by The Department of Health, Center for Health Statistics. The causes of death are determined by medical professionals. The dataset also contains the annual age-adjusted incidence rate of Lung

and Bronchus Cancer per 100,000 population, which is used to calculate the expected counts of mortality per county.

3 of the 39 counties had suppressed data (Garfield County, Wahkiakum County, and Columbia County). According to the WA department of health, this is likely due to the population falling below the threshold of 6 deaths (WA DOH, WTN Definitions). Additional WA state lung cancer rate among males was from the NIH State Cancer Profiles (https://statecancerprofiles.cancer.gov/incidencerates/index.php?stateFIPS=53&areatype=county&cancer=047&race=00&sex=1&age=001&stage=211&year=0&type=incd&sortVariableName=rate&sortOrder=default&output=0#results). This was used to generated expected counts for these 3 counties.

The data is available using WA state data portal: https://doh.wa.gov/data-and-statistical-reports/washington-tracking-network-wtn

Covariate Data

1. Personal Health Care Provider Access: Data with the counts, and percent, of adults who have reported to have at least one personal or primary health care provider. The data consists of results from the same year interval (2015-2019) as the cancer morbidity data used. This covariate is important in establishing whether possibly access to healthcare can ideally prevent cancer morbidity in regions through early screenings/treatments. The data is available through WTN: https://fortress.wa.gov/doh/wtn/WTNPortal/#!q0=1109
2. Median Household Income: To assess whether economic factors are also associated with lung cancer risk, median household income from 2015-2019 will be used. The data is available here: https://fortress.wa.gov/doh/wtn/WTNPortal/#!q0=481

Methodology

To calculate the expected cases of death from Lung & Bronchus Cancer in male populations, the Age Adjusted Rate per 100,000 was utilized, which was also provided by the Washington State Department of Health's WTN. The age-adjusted rate was derived from The National Cancer Institute's Surveillance Epidemiology and End Results (SEER). 4 model variations will be compared for mapping prevalence of male lung cancer mortality in Washington state, and evaluated for appropriateness.These models will be used to indicate counties that appear to have the highest risk for male lung cancer, for which resources and public health planning can allocate resources or possible interventions. The models include Standard Morbidity Ratio (SMR), IID models, spatial (BYM2) models, and spatial models with covariates.

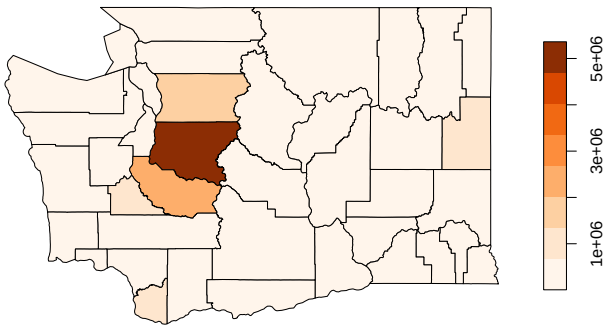
The IID Poisson-lognormal non-spatial random effect model used will be defined by:

$$\begin{aligned} Y_i|\beta_0, e_i &\sim_{ind} \text{Poisson}(E_i \mathrm{e}^{\beta_0} \mathrm{e}^{e_i}), \\ e_i|\sigma_e^2 &\sim_{iid} \text{N}(0, \sigma_e^2) \end{aligned}$$

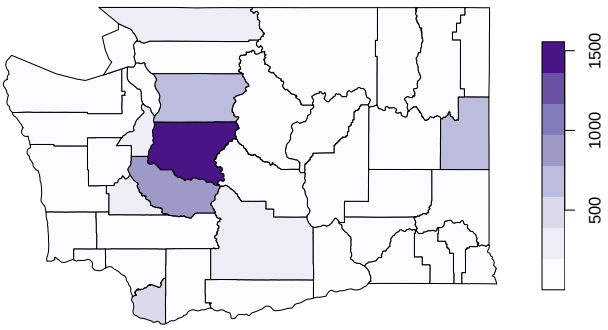
The Spatial Possion-lognormal ICAR random effects model is defined by:

$$\begin{aligned} Y_i|\theta_i &\sim \text{Poisson}(E_i\theta_i), \\ \log\theta_i &= \beta_0 + x_i\beta_1 + e_i + S_i, \\ e_i|\sigma_e^2 &\sim_{iid} N(0, \sigma_e^2) \end{aligned}$$

Statistical programming is done through `R` coding language, and supplementary files including the data files and code will be included with the project report. The `INLA` function will primarily be used for disease map modelling, with default priors utilized in the spatial and IID models. The covariate included be will analyzed to find any possible relationship to lung cancer risk, and likely to increase the predictive power of the model. To account for the three counties with suppressed data, the `cenpoisson` family was specified within `INLA`. This accounts for censored data when given the interval for possible values.



(a) Total Male Population



(b) Lung Cancer Mortality Count

Figure 1: WA County Maps, 2015-2019

Table 1: Counts

Male Population Total	Mortality Count Count
18225860	7412

Looking at the raw counts of mortality, and the population for WA state, we can see the highest numbers around specific areas of high population within the state. Particularly the Seattle Metro Area, which consists of King, Pierce, and Snohomish County is highly populated and located in close to the center of the state, west of the Cascade Mountains. Spokane County on the eastern border, and Clark county in the south east are also higher in population. Generally, Central and Eastern Washington are more rural. The Western border of the state from Pacific to Clallam county are also sparsely populated. The population and mortality count maps shown represent raw counts for population and cancer incidences, and do not take into account any form of spatial smoothing. This is important for sparse areas, where counts may be unreliable.

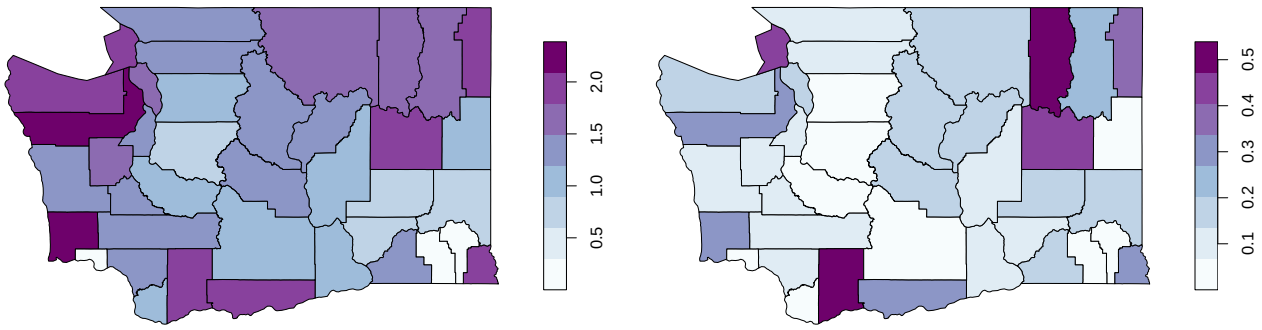
The response variable in this study would be lung & bronchus cancer deaths for males in WA state, and a Poisson distribution will be assumed for the disease. The disease is relatively rare compared to the entire combined population for the project (Table 1). A Poisson-Lognormal model in `INLA` will consequently be considered for this case, as a Poisson distribution is the most appropriate sampling distribution to assume for this study.

Results

Standard Morbidity Ratio (SMR)

The standardized morbidity ratio (SMR) is the ratio of deaths per county over the denominator group. In this case the denominator represents the male population per county in WA state, inclusive of all races, from 2015-2019.

To find the SMRs for each region, we can assume Y_i and E_i, $i = 1, ..., n$, denote the observed counts for the combined population in region i, $i = 1, ..., n$. Assuming $SMR = Y_i/E_i$, we can observe how this naive map would look with no spatial effects considered.



(a) SMR Map

(b) SMR Standard Errors

Figure 2: Lung Cancer Standard Morbidity Ratio Map

Table 2: 5 Counties with Highest SMRs

County	SMR	Standard Error
Jefferson	2.386194	0.3188686
Pacific	2.141632	0.2941758
San Juan	1.964083	0.4505914
Asotin	1.935774	0.3226289
Clallam	1.931260	0.1643998

Looking at the SMRs for Lung cancer in males (Fig 2A) shows contrast to the population and count maps (Fig 1). We can see that the Seattle Metro Area counties have generally lower SMR's, near or below 1 - indicating lower risk. The counties on the western border have relatively higher SMRs in contrast, with Jefferson county having the highest in the State. Most of central WA has low SMRs, with a slightly higher regions in the northeast. Klickitat & Skamania counties in the south also have high SMRs. Looking at the errors (Fig 2B), we can see the northeast regions have high errors, possibly due to being rural and sparse. Regions near King, Pierce and Snohomish county all have really small errors. The western boundary counties also have moderate errors. Table 2 shows the 10 tracts with the highest SMRs for Male Lung Cancer, and also the associated Standard Errors and County Names. Jefferson, Pacific, Clallam and San Juan counties all appear to have high SMRs and are located in the west boundary of the state. Smoothing can ideally lower the errors, and create a model that better reflects spatial relationships that these models don't.

IID Estimates:

Prior to any spatial smoothing, we can assume a simple IID model for INLA to observed what posterior medians for prevalence would be. This can be extracted using a Poisson-Lognormal model in INLA, and taking no additional covariates. The table below shows the Poisson Lognormal values for β and σ. The posterior median for the intercept is 0.28, with 95% intervals from 0.18-0.38. The total variance, σ, for the nonspatial model is 0.26.

Table 3: Poission Lognormal Non-Spatial

	Posterior Median	Lower CI	Upper CI
Beta	0.27	0.18	0.37

	Posterior Median	Lower CI	Upper CI
Sigma	0.26	0.19	0.35

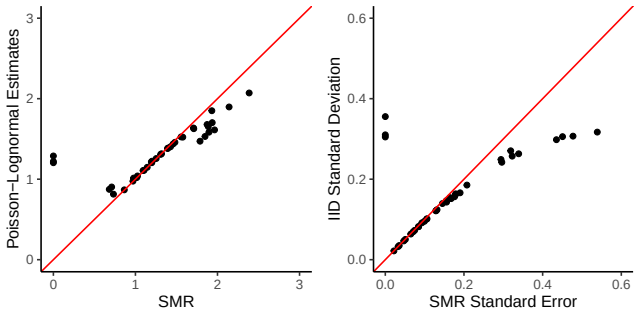


Figure 3: Comparison of Nonspatial Poisson-Lognormal model to SMRs

Just assuming a non-spatial IID model shows some contrast to the plot of Y_i counts for SMR. Supplementary Figures for the Poisson Lognormal non-spatial model shows similar risk hot spots near the northwest of the state, but also highlight the north east more. Regions in the center have slightly higher relative risks too. This can be confirmed by looking at a comparison of the IID model's relative risks and errors to the SMR's. Figure 5A shows shrinkage at the extremes of the SMRs, where low SMRs are slightly higher, and high SMRs are a lot smaller. Regions with SMRs between 1 and 1.5 are generally unchanged in the poisson-lognormal model. However, looking at the errors, we see that the poisson-lognormal model reduces the higher standard errors by a good amount. Regions with high initial errors (Ferry, San Juan, Skamania) see the largest reduction in errors. These were the counties in the South and Northeast, which this model is addressing better.

BYM2 Spatial Smoothing

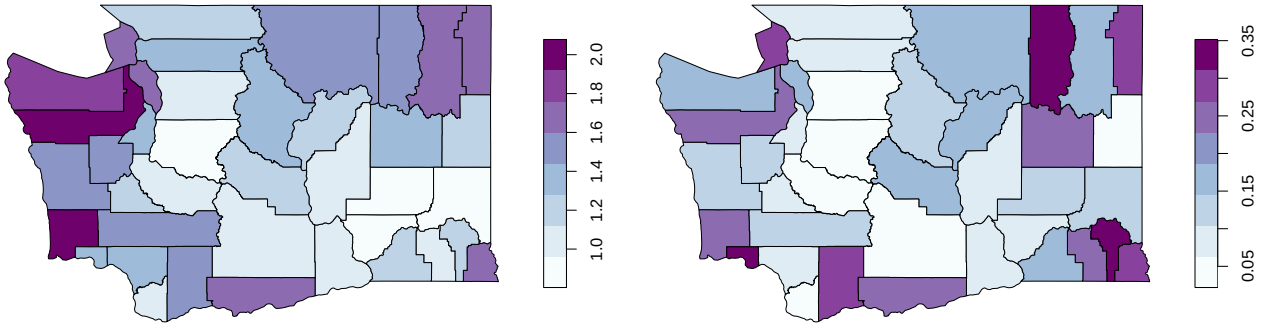
To leverage neighboring areas influence a spatial model will be looked at, that uses BYM2 spatial smoothing. The neighbor matrix for WA state census tracts is constructed using the `prioritizr` package in R. Default priors are assumed for this study. The spatial results (Table 4) shows that the β value (0.27) is similar to the non-spatial model (0.28), but with smaller confidence intervals. The total variance of random effects is 0.07, and a total variance 0.77 is coming from spatial random effects. The proportion of variance attributed to spatial effects has a wide 95% interval from 0.13 to 0.99.

Table 4: Spatial Poission Lognormal

	Posterior Median	Lower CI	Upper CI
Beta	0.26	0.20	0.33
Total Variance	0.07	0.04	0.13
Proportion Spatial	0.74	0.11	0.99

The effects of spatial smoothing is apparent in the North-West regions, where the relative risks are 'shared' between neighboring census tracts (Fig 4A). We see these regions have lower contrast in relative risks as a result. This makes the western boundary of the state hot-spots for higher prevalence of lung cancer, extending into the south-west counties such as Pacific, Lewis, Skamania, and Kickitat. The northeastern counties have slightly lower risks, borrowing from the central counties with really low estimates. So Pend-Oreille, Stevens, and Ferry have lower relative risks as a result. Figure 4B shows much lower errors than the SMR model, which is expected given smoothing.

Figure 5 can confirm these differences. We see most shrinkage for relative risks for high error counties (which generally also had high SMRs). At the lowest end, we see some increase in estimates too. Figure 6A shows



(a) Spatial Model Relative Risks

(b) Posterior Standard Deviations

Figure 4: Spatial (BYM2) Model

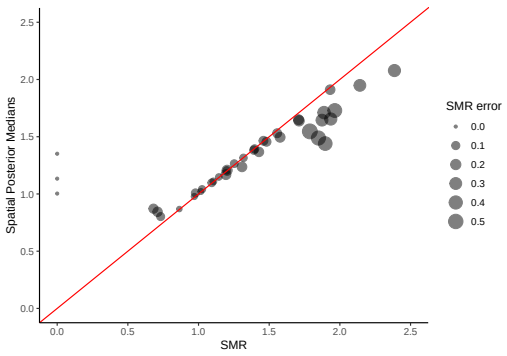


Figure 5: Comparing Spatial Model Risk to SMR

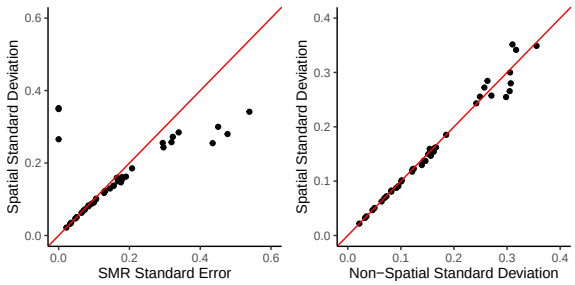


Figure 6: Comparing Spatial Model Relative Risks and Errors

reduction in errors, similar to the non-spatial model. When comparing the spatial poisson lognormal model to the non-spatial (Fig 6B), we see relative alignment, with differing values at the higher end of errors. Generally the spatial model slightly reduces errors for the western and southwest counties, whereas the non-spatial model has lower errors for the northeastern counties. Given the general high risk found along the western counties (Pacific, Jefferson, Clallam), I would suggest the spatial model would be more suitable.

Spatial Model with Covariates

The final model will consider the spatial poisson-lognormal model with the added covariates of median household income, and the access to a healthcare provider. We would generally expect a negative association for both these covariates, since income and access to healthcare could likely reduce the mortality rate from lung cancer.

Table 5: Spatial Model with Covariates

	Posterior Median	Lower CI	Upper CI
Beta	1.7642	0.9547	3.1918
Provider Access	1.0000	0.9999	1.0000
Median Household Income	0.9959	0.9853	1.0070
Total Variance	0.0501	0.0248	0.1009
Proportion Spatial	0.7073	0.1371	0.9830

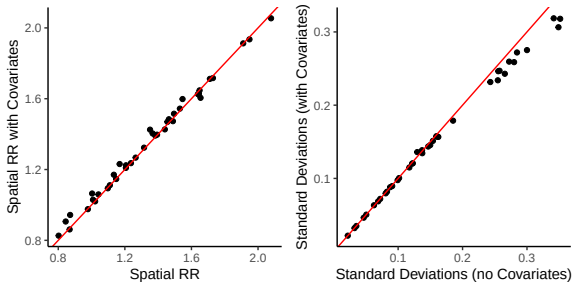
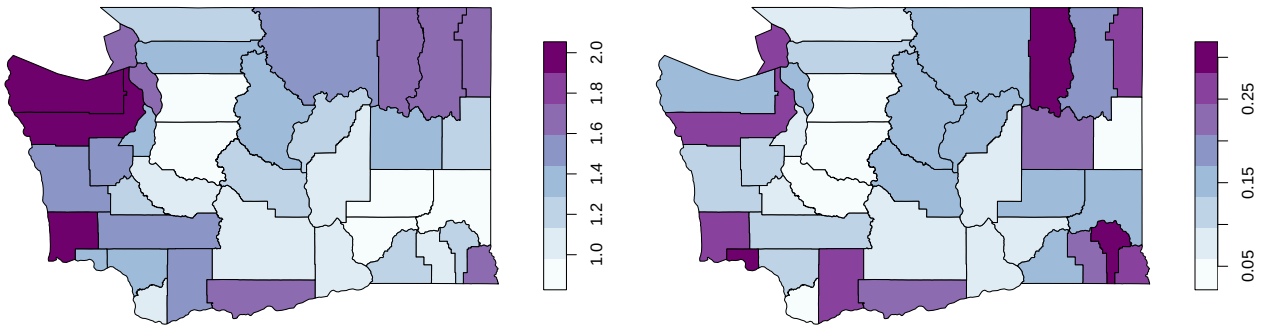


Figure 7: Comparing Spatial Model Relative Risks and Errors

Table 5 shows the spatial model with covariates of healthcare provider access and median income. Overall, there is not much of an association that can be made from these covariates given both have a relative risk median value of 1 (taking the exponentiated values of the log relative risks), and very small 95% confidence intervals. However there is an overall reduction in the total variance compared to Table 4 (BYM2 Spatial model), and proportion of variance attributed to spatial is also reduced slightly. We also see a slight reduction in the intervals. This implies that the model will likely have similar relative risk estimates compared to the spatial model lacking covariates. However it can be expected to see a slight reduction in error sizes.

Mapping the relative risks for counties using the spatial covariate model (Figure 8), there is alignment with most counties' prevalence to spatial model (Figure 4A). However, the estimates for Clallam and Ferry county appear slightly higher. Looking at Figure 7A, we can confirm similarities in Relative Risks between the two models, with slight differences on a county specific basis.

Overall Figure 8A shows spatial smoothing, maintaining higher prevalences in the eastern border of the state, and and northwestern corner. The Seattle metro area (where most of the population resides) maintains a low risk, as does most of the central and eastern counties. Figure 7B and 8B show that most county errors remain the same as the non-covariate model (Fig 4B). However, there is actually a reduction in errors in counties with spatial standard deviations of 0.2 or higher, when adding covariates (Fig 4B). This is most notable in the western counties of Jefferson and Pacific, and the northeastern counties of Ferry and Pend Oreille (Fig 8B).



(a) Spatial w/ Covariates, Relative Risks

(b) Posterior Standard Deviations

Figure 8: Spatial Model with Covariates

Discussion

Disease mapping was performed on lung cancer mortalities among males in Washington State. The data included incidences by county boundaries, between 2015-2019. The models considered for prevalence mapping included SMR, an IID model, a BYM2 Spatial model, and a Spatial model with 2 covariates. The chosen covariates were access to a primary healthcare provider and median income. Mapping the SMRs across all counties showed an initial trend where counties on the western border, particularly in the northwest, had highest SMRs. However, this also showed relatively high errors, particularly for the sparse rural areas near the northeast (which also showed a higher regional SMR). Notably, despite being the population center for Washington State, the Seattle surrounding metro area (King, Pierce, Snohomish, etc. . .) showed very low SMRs.

Applying a non-spatial IID model with no covariates saw a reduction in errors, while still maintaining relatively similar risk estimates with shrinkage at the extremes. When applying spatial smoothing though BYM2, we still observed agreement for the areas that generally had highest prevalence. However this model had the largest shrinkage in terms of relative risks, also also had lower errors than the SMR estimates. Mapping the spatial model, the counties with the highest prevalence for male lung cancer morbidity are Jefferson, Clallam, San Juan (Northwest Regions); Pacific (Western Border); Pend Oreille, Ferry, Stevens (Northeast Regions). These are all relatively sparse areas, and may be prime candidates for future public health studies and interventions to reduce the morbidity rates from Lung Cancer.

To understand potential associations for the spatial areas with high estimates, 2 covariates were included into the spatial model. Both covariates would address the readiness of populations to seek and access healthcare. One covariate was the number of respondents with healthcare access, and the other is the median income of the county. Running the model with the additional covariates showed weak associations, however the model did have a slight reduction in total variance and errors for the relative risks. Ultimately, it can be surmised that the chosen covariates did not indicate more information regarding the spatial nature of lung cancer morbidity among males in WA state.

Limitations

There are notable limitations to this study that must be addressed.

The Washington Tracking Network (WTN) provides lung cancer morbidity at a county level. This administrative level is rather large, with 39 total regions in the state. This may have cause finer trends to be lost to aggregation. For example, some of the large counties near the Salish Sea, such as Clallam county, are difficult to address. We don't know if the regional trend is throughout the entire county, or close to a specific subset of cities. A

finer administrative level analysis could have indicated trends within the state better, and possibly indicated whether this could be due to more specific factors (such as occupations in regions).

There are also limitations also to the chosen covariates. The two covariates chosen were to address readiness for medical care in the county, however there may have been better covariates to analyse such as occupations. Non-smoking related covariates were chosen since the connection to that has been well documented already, and so has air quality factors. There is a chance that the covariates were unresponsive given the large administrative levels too, so the relationship cannot confidently be approved or denied. The chosen demographic for this study was Male populations across all age groups and races. We can surmise that older individuals in this group may constitute higher death, so chosing a more specific age-group may have revealed stronger spatial associations with covariates.

I should also make note of the ecological bias that this study may have. Since I am working with aggregate data, it is easy to fall into the fallacy to attribute inferences of cancer prevalence to the region itself rather than the individuals that make up the aggregation of said data. It is can be well assumed that the chosen covariates have links to cancer morbidity. Generally having access to a healthcare provider, and higher income, would afford better screening and intervention opportunities. However, in this study, such a relationship was not apparent when looking spatially, however this does not negate their possible associations. Regions in the northwest of the state have higher lung cancer deaths, however this is an observation on general tendency for the aggregated counts for those counties, not a conclusion on the individuals living in those counties in particular.

Future Study

This project shows promising results for lung cancer studies in Washington State. There are many covariates that could be further studies, such as health expsoure factors which were not included. Other exposure factors that could also be looked into include wildfire smoke, which is particularly common in parts of eastern Washington. This study could also be done more focused on the Western half of the state that showed higher prevalence. For example, the high estimates for rural tracts away from the Seattle Metropolitan Area indicate some cause to higher cancer rates, however the reason is not clear from this study alone. I would suggest health intervention plans from public health specialists to better elucidate the relationship between known health access cofactors and lung cancer morbidity rates in the state.

References

1. https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html
2. Washington State Cancer Registry (October 2019). Cancer by Site Report. https://fortress.wa.gov/doh/wscr/Report.mvc/Report
3. Jemal A, Travis WD, Tarone RE, Travis L, Devesa SS. Lung cancer rates convergence in young men and women in the United States: analysis by birth cohort and histologic type. Int J Cancer. 2003 May 20;105(1):101-7. doi: 10.1002/ijc.11020. PMID: 12672038.
4. Washington DOH. (N/D). Definitions - Washington Tracking Network (WTN). doh.wa.gov. Available: https://doh.wa.gov/data-statistical-reports/washington-tracking-network-wtn/resources/definitions#S
5. NIH State Cancer Profiles. (Retrieved Mar 2023) Available at: https://statecancerprofiles.cancer.gov/incidencerates/index.php?stateFIPS=53&areatype=county&cancer=047&race=00&sex=1&age=001&stage=211&year=0&type=incd&sortVariableName=rate&sortOrder=default&output=0#results

Data Sources:

1. WA County Shapefiles: Washington State Department of Natural Resources (OpenData WADNR). (Retrieved Feb 2023). Available from: https://geo.wa.gov/datasets/wadnr::wa-county-boundaries/explore?location=47.18264 120.817600%2C7.57

2. Washington Tracking Network, Washington Department of Health. Web. "Cancer: Lung and Bronchus". Data were obtained from the Department of Health Center for Health Statistics, Community Health Assessment Tool (CHAT). Published June 2022. Available from: https://fortress.wa.gov/doh/wtn/WTNPortal/#!q0=1427