Areal Data Modeling: Lung Cancer Disease Mapping in Pennsylvania

Kayode Oyeniran

29/04/2021

1.0 Introduction

Cancer is a group of disease characterized by the uncontrolled growth and spread of abnormal cells that can results in death if not treated. Cancer varies by root or source, for instance, cancer that is rooted in the lung is referred as the lung cancer. Lung cancer is a type of cancer that develop when cells in the lung change. The change is triggered when people breathe in dangerous, toxic substance. Hence, the number one risk factor of lung cancer is "smoking." According to Center for Control and Disease Prevention (CDC), smoking accounts for about 90% of lung cancer cases worldwide. Other risk factors include radon, hazardous chemicals, particle pollution, gene etc. At the early stage, lung cancer can be attributed with symptoms such as incessant cough, constant chest pain, shortness of breath, pneumonia, coughing up blood etc., while at the later stage when lung cancer might have spread to other organs of the body, obvious symptoms like weight loss, loss of appetite, headaches, bone pain or fracture, blood clots etc., may show up.

Lung cancer is the leading cancer killer in both men and women in the United States in 1987, surpassing breast cancer to becoming the leading cause of cancer deaths in women. Thus far in 2021, lung cancer is the leading cancer killer in both male and female relative to other forms of cancer. According to cancer journal for clinicians, the estimated death rates in both male and female is 22% each. The age adjusted death rate for lung cancer is higher for men (46.7 per 100,000 persons) than for women (31.9 per 100,000 persons). It is similar for blacks (40.0 per 100,000 persons) and whites (39.2 per 100,000 persons) overall. Lung cancer is more prevalent in older people. In 2015, 86% of those living with lung cancer were 60 years of age or older. In 2018, there were 2.1 million new cases and 1.8 million death, making it the most common cancer worldwide. Hitherto in 2021, there are 235,760 new cases and 131,880 total death in the United States. In 2021 alone,

Pennsylvania has recorded 11,170 new cases of lung cancer, accounting for 13% of 85,440 total new cancer cases.

In this project, I have considered lung cancer incidence rate in Pennsylvania, United States, in the year 2002. Given the data, I developed a disease mapping model to estimate lung cancer risk in Pennsylvania. With the model in mind, I investigated a reasonable measure of the spatial association in the number of lung cancer cases in Pennsylvania by accounting for the proportion of smokers in each county.

2.0 Data

The dataset for this project is an areal data set from `SpatialEpi` package in R. The data set contains the population size, lung cancer cases, and smoking proportions for each of the counties in Pennsylvania in 2002. The lung cancer cases at the county level was stratified on race: a two level categorical variable (white and non-white), gender: a two level categorical variable (female and male), and age: a four level categorical variable (under 40, 40-59, 60-69 and 70+). For areal data modeling, the total number of cases per county was aggregated based on these strata. From the data, the expected number of cases was estimated for each county using the following formula:

$$E_i = \sum_j n_{ij} r_j$$

where r_j is the rate of lung cancer, indexed by age group, sex and race in the population of Pennsylvania,

n_{ij} is the population in each strata of county j.

Having defined the expected number of cases, I estimated the standardized morbidity ratios (SiRs) for each county. SMR is the ratio of observed cases to the expected cases, and it is a measure of risk of lung cancer in each of the counties in Pennsylvania.

2.1 Exploratory Data Analysis

Choropleth of Standardized Morbidity Ratios.

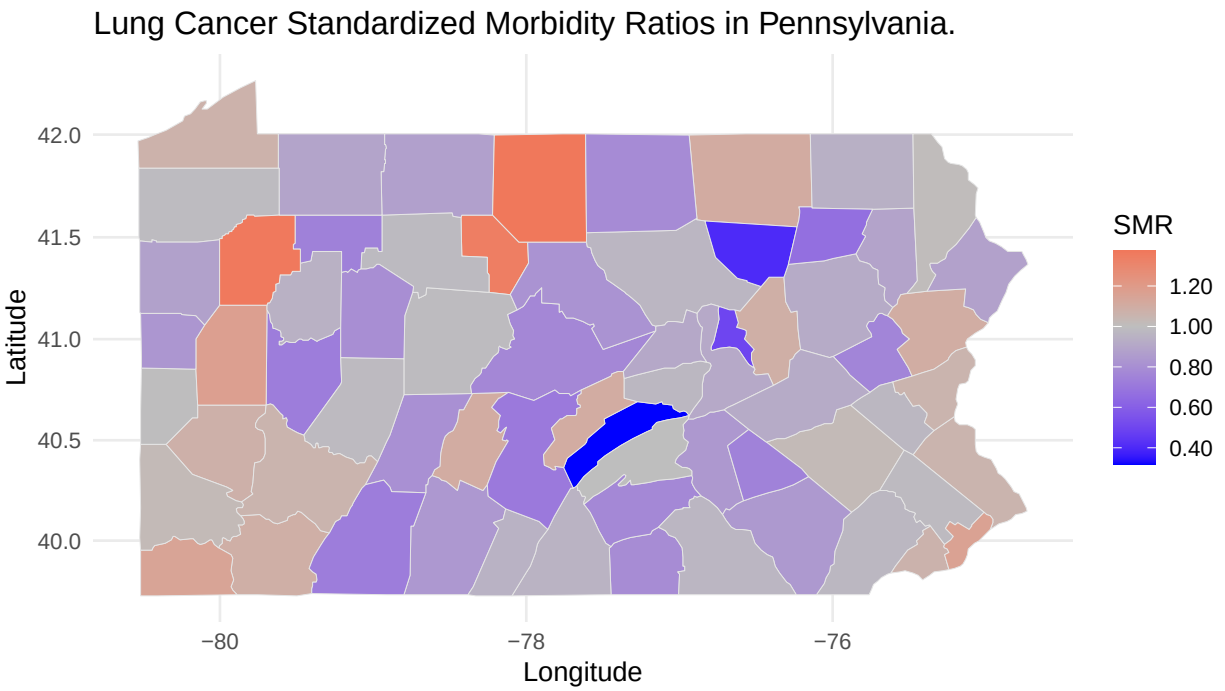


Figure 1 is the choropleth map, it helps us understand the risk of lung cancer across Pennsylvania. In counties with $SMR = 1$, the number of lung cancer cases observed is the same as the number of expected cases. In counties where $SIR > 1$, the number of lung cancer cases observed is higher than the expected cases. Counties where $SMR < 1$ have fewer lung cancer cases observed than expected.

GGmap of lung cancer cases.

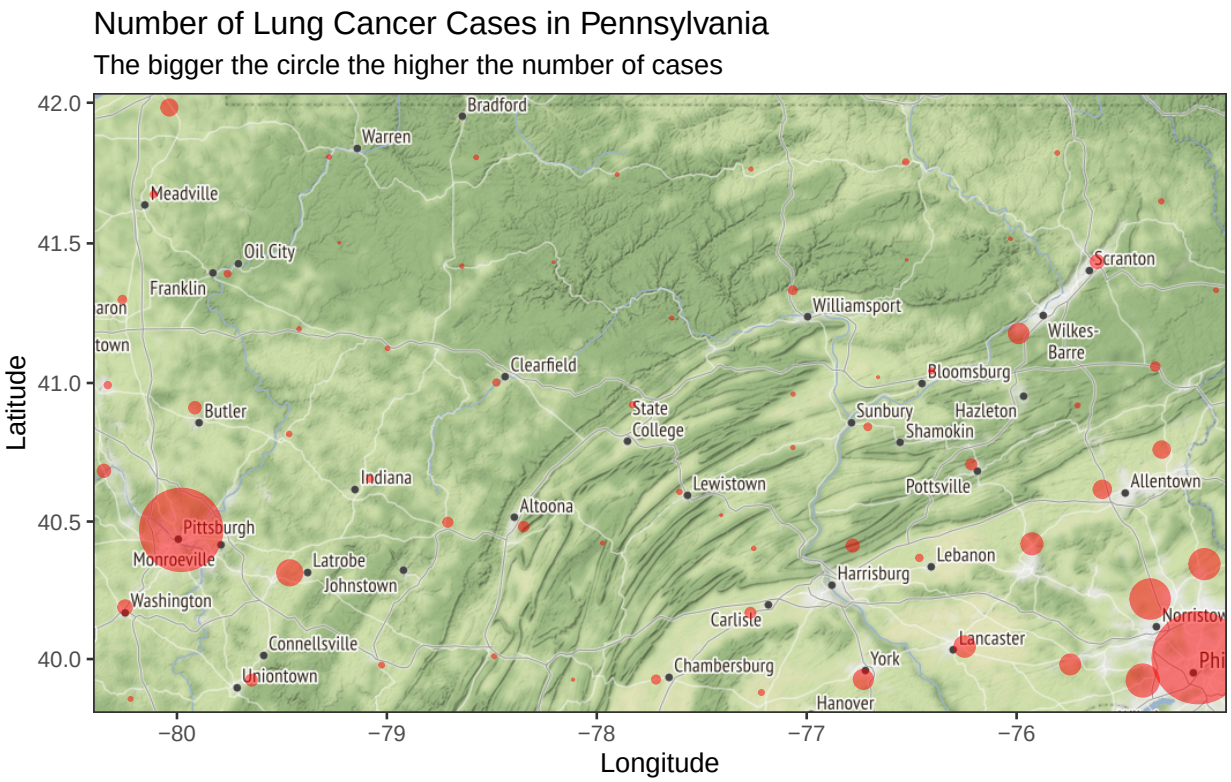


Figure 2 shows the size of lung cancer cases relative to other counties in Pennsylvania. Counties like Pittsburgh, Latrobe Johnson, Philadelphia, Delaware, Montgomery etc had prominent cases of lung cancer in Pennsylvania in 2002.

Histogram

Figure 3 shows the distribution of lung cancer cases across counties in Pennsylvania, and it appears to be heavily right skewed with two extreme values. Based on this figure, there appears to be a median lung cancer cases of 153.42 and inter-quartile range of 136 with associated minimum and maximum number of lung cancer of 3 and 1415, respectively.

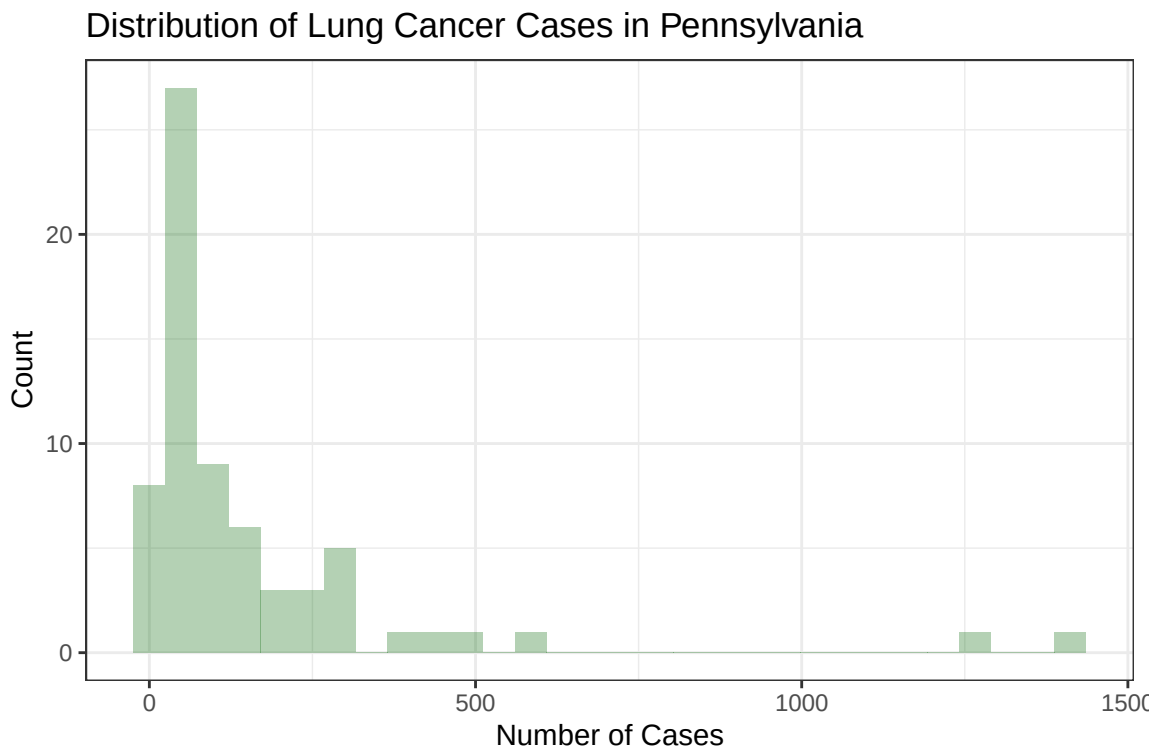


Figure 1: This figure shows the distribution of lung cancer cases in Pennsylvania.

3.0 Methods

For the model fit, we consider a model that accounts for the spatial associations between counties in Pennsylvania. This model allows us to borrow information from neighboring counties. Instead of conditioning the observed cases on the standardized morbidity ratios that fails to capture the disease risk in some counties due to population size differences, a Poisson model with spatial random effect is used that incorporate covariates information resulting in the smoothing of extreme values based on small sample sizes (Gelfand et al. 2010)

3.1 Model Specifications

Since the measurement of interest is the count of lung cancer (number of cases) in each county in Pennsylvania, then a Poisson model with spatial random effect will be appropriate to model the observed cases Y_i in county i.

$$
\begin{aligned}
Y_i|\psi_i &\sim Poisson(E_i \exp(\psi_i)) \\
\psi_i &= \boldsymbol{x_i^T \beta} + \theta_i + \phi_i
\end{aligned}
$$

Where E_i is the expected cases and $exp(\psi_i)$ is the relative risk of lung cancer in county i,

$X_i^T \beta$ is a linear combination of predictors or the mean structure,

X_i is the spatial covariate, and in this case, this is the proportion of smokers in each county,

θ_i is an ordinary random-effects components for non-spatial heterogeneity.

ϕ_i is the spatial random error.

4.0 Analysis

For the analysis, I used the log-normal Poisson model proposed by Besag York Mollié (BYM) 1991. The model includes both an CAR component for spatial smoothing and an ordinary random-effects component for non-spatial heterogeneity. The (BYM) model was adopted and the lung cancer disease risk estimates were obtained for each of the counties in Pennsylvania using the integrated nested laplace approximation (INLA) method.

Using INLA in R, the fitted model is given below:

- Model with with CAR spatial random effect and ordinary random effects for non-spatial heterogeneity.

$$
Log(\hat{\psi_i}) = -0.323 + 1.155 Smoking_i
$$

5.0 Model Summary

Based on our model, the intercept is -0.323 with a 95% credible interval of $(-0.621, -0.028)$. This is the estimated average lung cancer disease risk in county i with smoking proportion equal to 0. The coefficient of `smoking` (1.155) is the expected difference in number of lung cancer cases (on the logarithm scale) for one unit increase in smoking proportion with a 95% credible interval of $(-0.081, 2.384)$. In other hands, for one

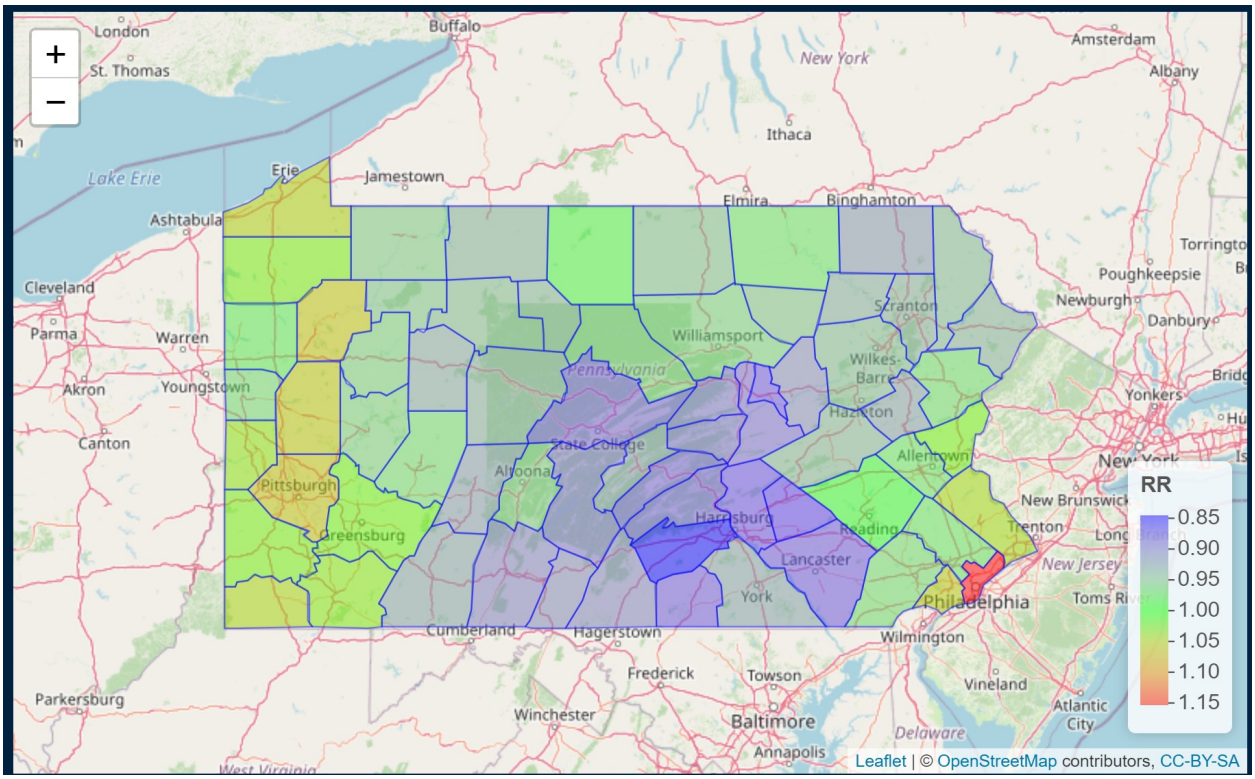
unit increase in smoking proportion, the incident of lung cancer increases by a factor of $exp(1.155)$, holding other predictor constant in the model.

Visualizing Risk of Lung Cancer Pennsylvania

To do this, I created a leaflet plot showing the risk of lung cancer for each of the counties in Pennsylvania in 2002. The mean risk estimates of lung cancer for each of the counties will be displayed on this plot, together with the standardized morbidity ratios, proportion of smokers, and the actual number of lung cancer cases. These risk estimates will be extracted from the model result from INLA. The posterior mean and corresponding 95% credible interval are in `model.fit$summary.fitted.values`.

Based on the leaflet map in Figure 4, it appears that majority of the counties with high risk of lung cancer are located in the west, with a few in the south east of Pennsylvania (specifically, Philadelphia, Bucks), and counties with lower risk of lung cancer are located in the center.. The corresponding 95% credible intervals indicate the uncertainty in the risk estimates.

Comparing the estimated relative risks of lung cancer in each of the counties with the standardized morbidity ratios, we observed some discrepancies in the two estimates. Based on prior knowledge, we know that SiRs can be misleading in population studies since it has the limitation of estimating correctly for counties or region with small population size. Using the Poisson log-normal model with a covariate structure together with the two random effect structures allowed us borrow information from neighboring counties to improve the risk estimates for each county, resulting in the smoothing of extreme values based on small population size.



Assessing Spatial Association

To assess the spatial association in the cases of lung cases in Pennsylvania, we fit separate model using the `S.CARleroux` function in R. Given our model, the estimated correlation coefficient ($\hat{\rho}$) of 0.901 shows strong evidence of positive spatial association. The choropleth in Figure 1 shows similar result using the SMR. We saw similarities in the estimated standardized morbidity ratios, particularly states that are close together. Hence, the evidence from the choropleth and the spatial model accounting for spatial structure provide strong evidence of a spatial association in the cases of lung cancer across counties in Pennsylvania in 2002.

Reference

Moraga, P. Small Area Disease Risk Estimation and Visualization Using R. The R Journal, 10(1):495-506, 2018 https://journal.r-project.org/archive/2018/RJ-2018-036/index.html

Moraga, P. (2019). Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny. Chapman & Hall/CRC Biostatistics Series, http://www.paulamoraga.com/book-geospatial

Gelfand, A. E., Diggle, P., Guttorp, P., & Fuentes, M. (Eds.). (2010). Handbook of spatial statistics. CRC press.

Appendix

1. BYM model

```
##
## Call:
##    c("inla(formula = formula, family = \"poisson\", data = map@data, ", "
##    E = exp.cases, control.predictor = list(compute = TRUE))" )
## Time used:
##     Pre = 0.524, Running = 1.08, Post = 0.428, Total = 2.04
## Fixed effects:
##               mean    sd 0.025quant 0.5quant 0.975quant   mode kld
## (Intercept) -0.323 0.150     -0.621   -0.323     -0.027 -0.323   0
## smoking      1.155 0.625     -0.083    1.157      2.386  1.160   0
##
## Random effects:
##   Name     Model
##     phi Besags ICAR model
##    theta IID model
##
## Model hyperparameters:
##                          mean       sd 0.025quant 0.5quant 0.975quant    mode
## Precision for phi     231.84   124.62      77.42   203.72     549.21  158.37
## Precision for theta 17533.84 17229.22    1088.97 12413.85   63111.43 2898.70
##
## Expected number of effective parameters(stdev): 18.71(4.49)
## Number of equivalent replicates : 3.58
##
## Marginal log-Likelihood:  -289.38
## Posterior marginals for the linear predictor and
##  the fitted values are computed
```

2. Spatial association assessment

```
##
```

```
## #################
## #### Model fitted
## #################
## Likelihood model - Poisson (log link function)
## Random effects model - Leroux CAR
## Regression equation - cases ~ smoking
## Number of missing observations - 0
##
## ############
## #### Results
## ############
## Posterior quantities and DIC
##
##             Median    2.5%  97.5% n.effective Geweke.diag
## (Intercept) 3.9060  2.2818 6.2118           9        -2.0
## smoking     1.6966 -7.9899 8.5086           9         2.0
## tau2        0.9262  0.6327 1.3937        1052        -1.0
## rho         0.9014  0.7224 0.9883         614        -0.7
##
## DIC =  557.966       p.d =  72.09461       LMPL =  -325.46
```