Evaluating the Relationship between

Lung Cancer and Radon Levels

in New York State

Jeanette Foxx

April 27, 2021

CHP 711

Dr. Galadima

**Abstract**

 It is well known that cigarette smoking is the leading cause of lung cancer. However, exposure to radon gas is the top cause of lung cancer in non-smokers, and the second leading cause of lung cancer overall (Stanifer, S., et. al., 2020). Many counties in New York State are at moderate to high risk of high radon levels. The purpose of this study is to determine if there is a significant relationship between lung cancer rate and radon levels in the state of New York. A multiple linear regression model was used to account for other causes and risk factors of lung cancer. There is a synergistic risk between smoking and radon exposure, meaning that exposure to both smoking and radon greatly increases the risk of lung cancer (Butler, K., et. al., 2018). Other predictors evaluated for the purpose of this study included age, sex, race, and occupational industry. Data were evaluated by county (n=58) to determine the significance of the association between lung cancer rate and the six predictor variables. The analysis showed a significant relationship between lung cancer rate and smoking, radon levels, and sex (p<0.05). These results indicate that exposure to high radon levels in certain counties of New York may be associated with the rate of lung cancer. Initiatives to educate the public of the risks posed by radon, especially in counties with high radon levels, are needed to encourage people to regularly test their home and take action if the levels are greater than the EPA action level of 4 picocuries per liter of air. Limitations of this study include limited radon testing data, as testing is not a regulatory requirement and is completely voluntary in personal residences. Radon test results for 15 out of 58 counties included less than 10 residences tested. Comprehensive testing is a recommended goal for counties with potential for moderate to high radon levels.

**Introduction**

 Lung cancer is the leading cause of cancer-related mortality and accounts for more than 1.7 million deaths globally every year (Brenner, H., et. al., 2020). Tobacco smoking is the top risk factor for lung cancer and is linked to nearly 90% of cases (Carter-Harris, L., et. al., 2017). There are other modifiable exposures associated with lung cancer such as secondhand smoke, radon exposure, and asbestos exposure. Lung cancer rates are also impacted by characteristics like sex, age, and place of residence. For example, it has been shown that incidence rates of lung cancer were higher in non-metropolitan counties than in metropolitan counties (O’Neil, M., et. al., 2019). Additionally, the rate of smoking is declining faster in metropolitan areas than in non-metropolitan. The incidence rate of lung cancer is higher in males than in females and higher in those greater than 55 years old (O’Neil, M., et. al., 2019).

 Exposure to radon is the second leading cause of lung cancer, resulting in an estimated 20,000 deaths each year in the United States (Butler, K., et. al., 2018). Radon is a colorless, odorless, radioactive gas that occurs naturally in the Earth’s crust (Stanifer, S., et. al., 2020). When this gas accumulates in indoor spaces, specifically a place of residence, it results in chronic radiation exposure to residents. It is recommended by both Federal and State government agencies that individuals test their homes for radon to determine if levels are elevated (Stanifer, S., et. al., 2020). The U.S. Environmental Protection Agency (EPA) recommends that radon levels greater than 4 picocuries per liter of air be quickly remediated to reduce exposure risk to the residents. Public awareness of the modifiable risk posed by radon remains low. One study in New York showed that even among residents who knew about radon, only 15% of them had their homes tested (Wang, et. al., 2000).

 This ecological study will examine the relationship between lung cancer incidence and average radon levels in counties of New York State. New York was chosen due to its geographical and population gradient, and due to the large amount of data available from the New York Health Department. In addition to radon levels, other variables analyzed include smoking rate, percentage of males, percentage of the population greater than 65, the percentage of the population that identifies as African American, and the percentage of the population that works in an industry which includes mining. Mine workers are at an increased risk of dying from lung cancer due to respiratory exposures in their occupation (Graber, J., et. al., 2014). They were included in this study as an additional predictor variable for lung cancer.

**Methods**

 The research will examine if there is a significant relationship between lung cancer rate and several predictors, including radon levels, in New York State. This ecological study will use a multiple linear regression model to analyze if there are significant correlations between the continuous response variable (lung cancer rate) and six explanatory variables. A multiple linear regression will be used for this study because there is one continuous response variable and there are six continuous predictor variables. The study consists of all counties in New York (n=58). New York City includes 5 boroughs that have been combined for the purpose of this study, including Bronx, Kings (Brooklyn), Queens, New York (Manhattan), and Richmond (Staten Island).

*Description of Variables*

 The dependent/ response variable (y) is lung cancer rate (incidence) in New York State over a 5-year period (2012-2016). Data was collected from the New York State (NYS) Cancer Registry and Cancer Statistics, as well as the Surveillance, Epidemiology, and End Results (SEER) Program Registry. The independent/ explanatory variables analyzed in this study include radon levels, smoking rate, percentage of males, percentage of African Americans, and percentage of the population that works in the agricultural and mining industry. Demographic data was gathered from the U.S. Census Bureau. Smoking rate in each county was gathered from the NY Department of Health’s Behavioral Risk Factor Surveillance System (BRFSS). Data for radon levels consisted of a 5-year average from radon tests taken in each country between 2012 and 2016. Results are reported in picocuries per liter of air (pCi/L) and were gathered from the NY Department of Health’s Environmental Public Health Tracker.

*Evaluation of Assumptions*

There may be a collinearity issue between radon levels and smoking status because there is a synergistic relationship between these variables and their effect on lung cancer rate (Lee. M., et. al., 1999). The variance inflation factor (VIF) will be used to measure independence between variables. Linearity and normality will be checked by evaluating the scatter plots and histograms using the SAS Analytics software program. Independence is assumed as the data were collected using well-designed sampling methods. Equal variances assumption will be checked by looking at the residual by predicted plot created by SAS and ensuring that the data points are randomly scattered above and below the X-axis.

*Testing the Hypothesis*

 The SAS software program will be used to run the linear regression model to test the following hypotheses*: Ho : b1 = b2 = b3 = b4 = b5 = b6 = 0 and Ha: Not all Bi = 0.* The F-value will be used to determine if the regression model is significant (if there in a linear relationship among the variables). If the null hypothesis is rejected, the slope will be evaluated to create a regression equation.

**Results**

*Checking Assumptions*

 Scatter plots shown in Appendix 1, Figures 1-6 display individual scatter plots of Y versus Xa to check for linearity. This assumption is met because each scatter plot shows a linear relationship. The independence assumption is met because data are independent and were collected using well-designed sampling methods. Figure 7 shows a histogram of the residuals and a normal distribution. Figure 8 shows the residuals tests. The residuals by percent plot shows a normal distribution. The residual by predicted value shows a random scatter of data points above and below the X-axis. Therefore, the normality and equal variance assumptions are met. Variance Inflation Factors (VIF) were checked for all variable to ensure no collinearity. No variable produced a VIF factor greater than 10, so collinearity is not considered an issue. All assumptions are met to allow for a multiple linear regression to be performed.

*Multiple Linear Regression Model*

 Table 1 shows the descriptive statistics for the data. There are no significant outliers evident in the data. The mean radon level was 4.76 pCi/L, and had a large range from 0.74-12.3 pCi/L. Other variables include mean percentage over 65 of 16.9%, mean smoking rate of 20.7%, mean percentage of males of 49.9%, mean percentage of African Americans of 5.2%, and mean percentage of workers in the mining industry of 2.3%.

| **Table 1: Descriptive Statistics of Variables** |
| --- |
| **Variable** | **Label** | **N** | **Mean** | **Minimum** | **Maximum** |
| RADONSMOKEMALERACEAGEINDUSTRY | RADONSMOKEMALERACEAGEINDUSTRY | 585858585858 | 4.757758620.713793149.88793105.165517216.85862072.2724138 | 0.74006.60047.7000.60012.2000.100 | 12.280030.50054.50022.0026.7008.100 |

 A multiple linear regression was first performed using all explanatory variables to determine which ones represent a significant correlation to lung cancer rate. Table 2 shows the parameter estimates of all explanatory variables and associated 95% confidence intervals. Smoking rate and percentage of males are the only variables to show a significant correlation with a p-value < 0.05. Attempts to obtain a better model fit were performed both manually and using automated model selection. Stepwise, forward, and backward selections were compared to determine best fit. Backward selection fit was not used because it only displayed one variable (age) as a significant predictor. Stepwise and forward models were identical and presented smoking rate, percentage of males, and radon levels as the three significant predictors of lung cancer. The forward model fit is shown in Table 3. A manual fit was performed by excluding insignificant variables one at a time until a best fit was achieved. The manual fit was identical to the best automated fit.

| **Table 2: Regression Model with All Variables and C.I.s** |
| --- |
| **Variable** | **Label** | **DF** | **ParameterEstimate** | **StandardError** | **t Value** | **Pr > |t|** | **95% Confidence Limits** |
| **Intercept** | Intercept | 1 | -92.91303 | 46.78125 | -1.99 | 0.0524 | -186.83030 | 1.00425 |
| **RADON** | RADON | 1 | 0.88070 | 0.44656 | 1.97 | 0.0540 | -0.01580 | 1.77720 |
| **SMOKE** | SMOKE | 1 | 0.47155 | 0.22986 | 2.05 | 0.0454 | 0.01009 | 0.93300 |
| **MALE** | MALE | 1 | 3.03181 | 0.91347 | 3.32 | 0.0017 | 1.19794 | 4.86567 |
| **RACE** | RACE | 1 | -0.64530 | 0.36783 | -1.75 | 0.0854 | -1.38376 | 0.09316 |
| **AGE** | AGE | 1 | 0.39323 | 0.53713 | 0.73 | 0.4675 | -0.68511 | 1.47157 |
| **INDUSTRY** | INDUSTRY | 1 | -1.66381 | 0.77955 | -2.13 | 0.0376 | -3.22883 | -0.09879 |

| **Table 3: Forward Selection Model with Outliers** |
| --- |
| **Variable** | **Label** | **DF** | **ParameterEstimate** | **StandardError** | **t Value** | **Pr > |t|** | **VarianceInflation** |
| **Intercept** | Intercept | 1 | -80.91827 | 43.60590 | -1.86 | 0.0690 | 0 |
| **RADON** | RADON | 1 | 1.01241 | 0.45345 | 2.23 | 0.0297 | 1.06290 |
| **SMOKE** | SMOKE | 1 | 0.60924 | 0.21541 | 2.83 | 0.0066 | 1.14107 |
| **MALE** | MALE | 1 | 2.71192 | 0.88987 | 3.05 | 0.0036 | 1.09016 |

Residual diagnostics identified three potential outliers at points 4, 16, and 57. The outliers could not be attributed to data entry errors and, therefore, were omitted one at a time to look at the effect on the fit. Table 4 shows the model fit with outliers removed.

| **Table 4: Forward Selection Model with Outliers Removed** |
| --- |
| **Variable** | **Label** | **DF** | **ParameterEstimate** | **StandardError** | **t Value** | **Pr > |t|** | **VarianceInflation** |
| **Intercept** | Intercept | 1 | -162.01071 | 49.81549 | -3.25 | 0.0020 | 0 |
| **RADON** | RADON | 1 | 0.99727 | 0.44713 | 2.23 | 0.0302 | 1.05918 |
| **SMOKE** | SMOKE | 1 | 0.72094 | 0.20277 | 3.56 | 0.0008 | 1.08189 |
| **MALE** | MALE | 1 | 4.31291 | 1.01261 | 4.26 | <.0001 | 1.03128 |

The fit was improved by removing the outliers, which is evident in the ANOVA table and associated R2 value (shown in Table 5). The R2 value and F coefficient increased with outliers removed, indicating increased significance of the model. The p-value of the F-statistic is <0.0001, showing that we can reject the null hypothesis because there are explanatory variables that are significant predictors of lung cancer rate. The R2 number is 0.4881, meaning that nearly 49% of the total variation in the response variable is explained by the regression model. The parameter estimates provided in Table 4 were used to create the multiple linear regression equation for this fit:

$$\hat{y } = -162.011+0.997X1+0.721X2+4.313X3 $$

| **Table 5: ANOVA Table of Final Fit Selection** |
| --- |
| **Source** | **DF** | **Sum ofSquares** | **MeanSquare** | **F Value** | **Pr > F** |
| **Model** | 3 | 3309.85541 | 1103.28514 | 16.21 | <.0001 |
| **Error** | 51 | 3470.72096 | 68.05335 |  |  |
| **Corrected Total** | 54 | 6780.57636 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 8.24945 | **R-Square** | 0.4881 |
| **Dependent Mean** | 71.85455 | **Adj R-Sq** | 0.4580 |
| **Coeff Var** | 11.48076 |  |  |

**Conclusion**

 This study examined the relationship between lung cancer and several predictor variables including radon levels, smoking rate, sex, race, age, and industry. The best fit for this multiple linear regression model suggests that radon levels, smoking, and being a male may be significant predictors of lung cancer. One study performed in New York showed that lung cancer incidence was highest among males who lived in non-metropolitan areas within the state (O’Neil, M., et. al., 2019). The shaded areas in Figure 9 in Appendix 1 show counties of New York with elevated radon levels (> 4 pCi/L). As visible by this map, over half of the state contains residents with levels that are considered unsafe. Based on the data used in this study, the top five counties with the highest radon levels included Oneida, Cattaragus, Greene, Cortland, and Jefferson with an average radon level of 10.7 pCi/L, which is over two times higher than the EPA action level of 4 pCi/L. New York City presented the lowest lung cancer incidence, as well as the lowest radon levels and lowest smoking rate. Over the past several decades lung cancer due to smoking has steadily declined, preventing over 800,000 cancer deaths it is estimated (2015). Still, other causes of lung cancer, such as radon exposure, require awareness. Lung cancer caused by radon is highly preventable by regular testing and mitigation. One study showed that over half of participants conducted radon testing within their homes when presented with a free testing kit (Butler, K., et. al., 2018). Performing the test is simple and passive on the part of the resident. Making tests free and available is an effective way to increase the level of testing, and a way to increase knowledge in the population. Another initiative includes proposed legislation to ensure radon testing and mitigation, if necessary, during real estate transactions (Stanifer, S., et. al., 2020). This bill would be effective to ensure homebuyers are aware of the risks of radon and homes with elevated levels are mitigated prior to new resident occupancy. Other options include building new construction with radon-resistant materials to ensure radon cannot enter and accumulate within the home.

*Study Limitations*

 Limitations to this model include a lack of radon testing data. There were 15 of the 58 counties analyzed that had less than 10 residences tested. Therefore, the average radon levels for those counties may be misleading. Testing of more homes may cause the average radon level to increase or decrease. Comprehensive testing is a recommended to further this study and as a goal for counties with potential for moderate to high radon levels.

**References**

Brenner, H., & Krilaviciute, A. (2020). Commonly Applied Selection Criteria for Lung Cancer Screening May Have Strongly Varying Diagnostic Performance in Different Countries. *Cancers, 12*(10), 3012.

Butler, K. M. (2018). Access to Free Home Test Kits for Radon and Secondhand Smoke to Reduce Environmental Risks for Lung Cancer. *Journal of Environmental Health, 81*(3), E1-E6.

Carter‐Harris, L., Ceppa, D. P., Hanna, N., & Rawl, S. M. (2017). Lung cancer screening: what do long-term smokers know and believe? *Health Expectations, 20*(1), 59-68.

Graber, J. M., Stayner, L. T., Cohen, R. A., Conroy, L. M., & Attfield, M. D. (2014). Respiratory disease mortality among US coal miners; results after 37 years of follow-up. *Occupational and Environmental Medicine, 71*(1), 30-39.

Lee ME, Lichtenstein E, Andrews JA, Glasgow RE, Hampson SE. Radon-smoking synergy: A population-based behavioral risk reduction approach. *Preventive Medicine. 1999 Sep;29(3):222-7*.

O'Neil, M. E., Henley, S. J., Rohan, E. A., Ellington, T. D., & Gallaway, M. S. (2019). Lung Cancer Incidence in Nonmetropolitan and Metropolitan Counties - United States, 2007-2016. *MMWR: Morbidity & Mortality Weekly Report, 68*(44), 993-998.

Stanifer, S., & Hahn, E. J. (2020). Analysis of Radon Awareness and Disclosure Policy in Kentucky: Applying Kingdon's Multiple Streams Framework. *Policy, Politics & Nursing Practice, 21*(3), 132-139.

Wang, Y., Ju, C., Stark, A. D., & Teresi, N. (2000). Radon awareness, testing, and remediation survey among New York State residents. *Health Phys, 78*(6), 641-647.

U.S. Department of Health and Human Services. (2015, October 1). *Nearly 800,000 deaths prevented due to declines in smoking*. National Institutes of Health. https://www.nih.gov/news-events/news-releases/nearly-800000-deaths-prevented-due-declines-smoking.

**Appendix 1: Figures and Tables**

Figure 1: Scatter Plot Cancer v. Radon

Figure 2: Scatter Plot Cancer v. Smoking



Figure 4: Scatter Plot Cancer v. Industry

Figure 3: Scatter Plot Cancer v. Sex



Figure 6: Scatter Plot Cancer v. Race

Figure 5: Scatter Plot Cancer v. Age



**Figure 8: Equal Variance Check, Residual by Predicted Value**

**Figure 7: Normality Check, Histogram of Residuals**



Figure 9: NY Counties with High Radon Levels



**Appendix 2: SAS Codes**

*Scatter Plots*

proc sgplot data = work.dap;
reg x = radon y = cancer;
title "Scatter Plot with Linear Fit";
xaxis label = "Radon Level";
yaxis label = "Lung Cancer Rate";
run;

proc sgplot data = work.dap;
reg x = smoke y = cancer;
title "Scatter Plot with Linear Fit";
xaxis label = "Smoker Percentage";
yaxis label = "Lung Cancer Rate";
run;

proc sgplot data = work.dap;
reg x = male y = cancer;
title "Scatter Plot with Linear Fit";
xaxis label = "Percentage Male";
yaxis label = "Lung Cancer Rate";
run;

proc sgplot data = work.dap;
reg x = race y = cancer;
title "Scatter Plot with Linear Fit";
xaxis label = "Percentage Black Race";
yaxis label = "Lung Cancer Rate";
run;

proc sgplot data = work.dap;
reg x = industry y = cancer;
title "Scatter Plot with Linear Fit";
xaxis label = "Mining Industry Percentage";
yaxis label = "Lung Cancer Rate";
run;

proc sgplot data = work.dap;
reg x = age y = cancer;
title "Scatter Plot with Linear Fit";
xaxis label = "Percentage Over 65";
yaxis label = "Lung Cancer Rate";

*Multiple Linear Regression*

proc reg data=work.dap;
model cancer = radon smoke male race age industry;
test radon;
Run;
quit;

\*forward\*;
proc reg data=work.dap;
model cancer = radon smoke male race age industry/ selection= forward slentry =0.05 vif r;
run;
quit;

\*backward\*;
proc reg data=work.dap;
model cancer = radon smoke male race age industry/ selection= backward slentry =0.05 vif r;
run;
quit;

\*stepwise\*;
proc reg data=work.dap plots(label) = (CooksD RStudentByLeverage);
model cancer = radon smoke male race age industry/ selection= stepwise slentry =0.05 vif r;
run;
quit

*Residuals Diagnostics*

Proc reg data=work.dap plots(label) = (CooksD RStudentByLeverage);

model cancer = radon smoke male;

run;

quit;