Predicting Cancer Rates Based on Radon and Smoking Levels Using Machine Learning

Tony Gwyn, BS1, Wendell Futrell, BS1, Umit Topaloglu, MD2
1Winston Salem State University, Winston Salem, NC and 2Wake Forest School of Medicine, Winston Salem, NC

Introduction

Cancer is the leading cause of death in North Carolina, having surpassed heart disease in 2009. Lung Cancer in particular exceeds the national average in our state. Smoking is both a major risk factor for Lung Cancer and COPD. Slightly less known is that Radon exposure can also quite dramatically increase the risk for both diagnoses. Studies have already shown that smoking and Radon exposure have a dramatic effect on cancer rates. As such we propose using machine learning to predict cancer probability based on a patient’s smoking and Radon risk factors.

Methods

Using a dataset of 6500 patients provided by Wake Forest Baptist Health, a database was created. Around 5500 patients had a positive lung cancer or COPD diagnosis, while the rest presented with no lung cancer or COPD. The data was eventually pared down to just over 6300 patients, as some patient data was incomplete or missing. Statistically significant fields included location by zip code, smoking rates, Radon levels, age, and overall diagnosis. The main programming language used to work with this data was Python, and the major packages used to synthesize the data was TensorFlow, Pandas, and Numpy.

Smoking rates were split into three categories and given weighted values. A non-smoker received a weighted value of 1, a former smoker a value of 2, and a current smoker a value of 3. Similarly, the Radon data was split into three categories. Radon data for this study was measured in picoCuries per liter of air, and measured on a by zip code basis. For a picoCurie level between 0 and 2, the weighted value was 1. Between 2 and 4 picoCuries received a value of 2, and 4 or more picoCuries received a value of 3. The overall diagnosis was split into two categories; one for smoking/Radon related diagnosis, and the other for non-smoking/Radon related diagnosis. A logistic regression based on the data was created, using 89 percent of the entire data to train the model, and the remaining 11 percent to test the data.

Results

The results of this study showed that as smoking levels increased, so did the occurrence of a positive lung cancer or COPD diagnosis, which is in line with other similar studies. The Radon data was a bit more mixed, showing that the highest occurrence of positive diagnosis happened in the 2 to 4 picoCurie range, while the 4 and higher picoCurie range lagged just behind. The amount of data points in the 2 to 4 picoCurie range far outpaced those in the 4 or higher range; having more data from patients in a higher Radon zip code would likely change the results.

As the amount of positive diagnosis far exceeded the amount of negative diagnosis in the data, that data was adjusted to give equal populations for each diagnosis using a function available in TensorFlow. The resultant receiver operating characteristic curve showed favorable results. The precision of the data was 87 percent, while the recall and the accuracy were both 77 percent. The area under the curve for the ROC was a .79.

Conclusion

Knowledge is power, and the knowledge that smoking, as well as Radon exposure, greatly increases one’s risk of developing lung cancer or COPD in the future might help them to move to mitigate those factors. The logistic regression created in this study proved that it is possible to accurately estimate a person’s risk of having a lung cancer or COPD diagnosis based on their smoking, Radon, and other factors. A larger dataset, as well as the addition of other factors that were not considered in this study will likely improve the functionality of this logistic regression and increase the accuracy in the future.