**Pesticide Exposures and Lung Cancer**

Does the overall cumulative exposure to various airborne pollutants increase the risk of lung cancer for smokers and non-smokers?

The World Health Organization names lung cancer the most common cause of cancer deaths in 2020 with over 1.8 million losing the battle. While cigarette smoking is a known contributing factor to the development of lung cancer, less attention has been given to other types of air exposures. Here we will explore the possible relationship between total exposures to a variety of airborne exposures by visualizing the dataset

‘Pesticide and lung cancer’, juntarawijit, chudchawal (2014-2017); <https://doi.org/10.6084/m9.figshare.12356270.v5>.

This data was previously analyzed presented in research paper: Pesticide exposure and lung cancer risk: A case-control study in Nakhon Sawan, Thailand : Teera Kangkhetkron and Chudchawal Juntarawijit (2020)

The dataset contains 68 columns which record responses of participants in Nakhon Sawan province, Thailand by in-person questionnaire. 233 lung cancer cases, and 458 healthy neighbors matched for gender, and age (±5 years) were questioned about pesticides and other types of environmental exposures. Demographic data, distance to nearest farm land, air pollution exposures, cigarette smoking habits were collected. Many of the features were recorded in multiple forms such as yes or no, number of days, total per year, etc. Many cells were left empty when the field is a quantity following a no exposure response.

Cigarette smoking was recorded in 5 different columns. Here we will look at Cigarette\_total and CigSmoke\_status.  
Data Dictionary: <https://figshare.com/articles/dataset/Pesticide_and_lung_cancer/12356270/5?file=39615298>

Target: ID: Float data. .0 as control, no lung cancer and .(1-9) as lung cancer

LungCA: Lung cancer status of responders: 0 as control and 1 as having lung cancer

Cigarette\_number: the number of cigarettes smoked per day

CigSmoke\_status: Tobacco status of responders, 1 refers those who have never smoked a cigarette, 2 refers to ex-smoker, 3 refers to current smoker

age\_group: Responders in each group;

1 refers to those with age less than or equal to 54, 2 refer to those with age 55-64 yr, 3 refer to those with age 65-74 yr, 4 refer to those with age 75 yr or more

We will start by exploring the dataset for missing values, unclear labels, dataframe types to determine cleaning and munging steps using a Python Jupyter notebook and the Pandas library.

lung\_cancer.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 680 entries, 0 to 679

Data columns (total 68 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 ID 680 non-null float64

1 LungCA 680 non-null int64

2 Gender 680 non-null int64

……………………………………………………………………………………………………………………………………………………………………………………..

67 Morphology\_Group 680 non-null int64

dtypes: float64(1), int64(43), object(24)

memory usage: 361.4+ KB

We can see there are 68 columns with 680 rows of data corresponding to the answers for each respondent in the survey. The dataset contains float, integer and object data types and no row appears to be empty or contain null or NaN characters.

This is a quite a wide dataset which would not fully display using default settings in Jupyter. Pandas display options where added to set the display dimensions, column width and the maximum number of columns to display. These settings allow us to view all the columns by adding a slider in the result window. A graphic of all of the columns would not be practical here so I’ve included only a small section chosen using the df.sample() command.

A screenshot of a computer

Description automatically generated

Note the empty cells in the above snippet. These cells went under the radar of our previous attempts to uncover missing or empty cells! Sneaky devils! The dataset was loaded from a CSV file for which Pandas interprets empty cells as containing white space and is a data. Regular expressions (Regex) is a good method to replace the white space with a value of interest. Since all of the empty cells follows a response of 0 or ‘no’ to any use of the specified pesticide or pollutant, these empty cells can safely be replaced with a value of 0. It follows that no use of a pesticide would have zero days of use and zero years of use.

lung\_cancer = lung\_cancer.replace(r'^\s\*$', 0, regex=True)

A screenshot of a computer

Description automatically generated

Problem solved! All of the empty cells are not ‘0’.

Now for a bit of munging before digging in to visualize our dataset. 68 columns is a lot of data to visualize. Let’s bring that down. Much of this dataset has been recorded in repetition using slightly different quantifier for each column. For example, a yes or no response followed by a column for how many days of use. Cigarette smoking is separated into six different columns beginning with each column providing more detail. Cigarette status containing never smoked, former smoker and current smoker as well as total life time number of cigarettes smoked are of greatest interest. As are the columns containing the total number of exposures. A new dataframe, lung\_cancer\_trunc, with only the columns of interest is created to make the dataset more manageable.

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 680 entries, 0 to 679

Data columns (total 30 columns):

30 columns, this should be more manageable. Just a little more munging to clarify some of the columns labels, change the datatypes to integer for graphing and calculations and map words to some of the integer coded responses to provide clear graph labels.

lung\_cancer\_trunc=lung\_cancer\_trunc.astype(int)

lung\_cancer\_trunc['ID']=lung\_cancer['ID']

lung\_cancer\_trunc['LungCA'] = lung\_cancer\_trunc['LungCA'].map({1: 'Yes', 0: 'No'})

lung\_cancer\_trunc['CigSmoke\_status'] = lung\_cancer\_trunc['CigSmoke\_status'].map({1: 'Never', 2: 'Quit', 3: 'Current'})

lung\_cancer\_trunc['Lung Cancer'] = lung\_cancer\_trunc['LungCA']

Beautiful! On to the fun part. Visualization of the data of interest!

A quick check to look at total cigarette smoking’s contribution to lung cancer development against age groups from younger to older age groups.

<seaborn.axisgrid.FacetGrid at 0x1eefde39550>

A graph showing the growth of cancer

Description automatically generated

Pretty telling but no new discovery here. This relationship is well established.

Creating relational plots of each individual pesticide days of use by age groups split by smoking status with lung cancer outcome as the line hue will provide a detailed visual of cancer development by increased exposure. Below are a few of the more significant graphs.

A computer code with red and blue text

Description automatically generated

A graph of a graph of a person

Description automatically generated with medium confidence

A graph of a graph

Description automatically generated

There may be an increased incident of lung cancer for smokers and non-smokers of all ages with a larger number of exposures to Chloropylifos or Carbofuran based substances. Chlorpylifos is an insecticide which is harmful to honey bees. Carbofuran is also an insecticide which is used throughout the world on crops.

A graph of a number of people

Description automatically generated with medium confidence

Folidol is a highly toxic insecticide developed in the 1940’s which has been banned in most countries. This insecticide might contribute more to younger non-smoker’s lung cancer and cause additional lung damage for those currently smoking.

These graphs show isolated exposures but it many of the responders used a number of pesticides over the years. Total exposures would be a better way to visualize an overall effect. Let’s create a new column by multiplying the days of use by number of year of the three categories of pesticides and add them together then graph by age group and cancer status. We’ll exclude cigarette smoking for now.

lung\_cancer\_trunc['Total\_exposures'] = (lung\_cancer\_trunc['Herbicide\_day']\*lung\_cancer\_trunc['Herbicides\_year'])+(lung\_cancer\_trunc['Number\_Days\_Insecticides\_Use\_432']\*lung\_cancer\_trunc['Number\_Years\_Insecticides\_Use\_Group\_432'])+(lung\_cancer\_trunc['Number\_Days\_Fungicides\_Use\_433']\*lung\_cancer\_trunc['Number\_Years\_Fungticides\_Use\_433'])

A graph of cancer and lung cancer

Description automatically generated

Interestingly, there does appear to be a relationship between the use of pesticides and lung cancer for all ages groups but more so for the younger and older adults.

Again, we’ll break this down by smoking status to see if we can discover if cigarette smoking is contributing to the increase in lung cancer we see here.

A close-up of a number

Description automatically generatedA graph showing the number of patients

Description automatically generated

A close-up of a number

Description automatically generatedA graph with a line

Description automatically generated

A close-up of a number

Description automatically generatedA graph with lines and numbers

Description automatically generated

Pesticide use does appear to have a causative effect increasing with the person’s smoking status.

Visualizing the dataset gives us an idea of what relationships might be present. However, drawing conclusions here could be very misleading. Although the graphs appear to indicate a relationship, it cannot tell us how strong the relationship might be. For this we need to use statistical tests. This may reveal a strong or weak relationship and help us understand how likely or unlikely the relationship would be due to random chance.

Pearson’s r correlation coefficient measures the linear relationship between two variables. It ranges from -1 to 1 with -1 and 1 being perfectly correlated. 1 meaning as one goes up so does the other and -1 being as one goes up the other goes down. Applying this static here may prove quite useful.

Carbofuran exposure:

scipy.stats.spearmanr(lung\_cancer\_trunc['Lung Cancer'], lung\_cancer\_trunc['Carbofuran\_days'])

SignificanceResult(statistic=0.142542298547309, pvalue=0.00019203626445637946)

The relationship between exposure and Carbofuran are highly unlikely to have occurred by chance but the relationship is not highly correlated. There are quite possibly other unknown confounding factors in play. But given that the relationship is unlikely to be caused by chance, it would be worth further research to determine if the link exists and how strongly.

Total exposures:

scipy.stats.spearmanr(lung\_cancer\_trunc['Lung Cancer'], lung\_cancer\_trunc['Total\_exposures'])

SignificanceResult(statistic=0.14156920288998381, pvalue=0.00021260352689352128)

Nearly identical to the Carbofuran exposure results. Perhaps Carbofuran is the greatest contributor out of all the types of pesticides. Again, we cannot determine this from the dataset.

Finally, let’s create another new column combining total exposures with the number of cigarettes smoked in a lifetime and again graph against lung cancer.

lung\_cancer\_trunc['Cig\_plus\_exposures']=lung\_cancer\_trunc['Total\_exposures']+lung\_cancer\_trunc['Cigarette\_total']

A graph of different colored squares

Description automatically generated

Cumulative exposures again appear to increase cancer development. Let’s see if spearman’s r looks any different now.

scipy.stats.spearmanr(lung\_cancer\_trunc['Cig\_plus\_exposures'], lung\_cancer\_trunc['Lung Cancer'])

SignificanceResult(statistic=0.11856238326372222, pvalue=0.0019550938054608055)

Though unlikely to have occurred by chance, total combined exposures is not a strong linear relationship with lung cancer development. Further research would be useful to determine if a risk exists and for whom.