Cancer Death Rates Based on Total Hospital Admissions & External Factors

Amit Das (akdas), Pratap Gude (gvnsp), Eman Wong (emanwong)

SI 618

April 12, 2023.

GitHub Link

Introduction & Motivation

Cancer has been a leading cause of mortality in humans and animals for centuries, with limited or only palliative treatment options. But what are the factors that contribute to high cancer rates? How do these factors change over time, and are they related to our everyday lives?

Despite decades of research, Cancer pathogenesis is highly unpredictable and challenging, as it has multifactorial causation factors involved in cancer etiology, prognosis, and recurrence. Various genetic, environmental, lifestyle and other factors contribute to the etiology and progression of the condition. During our background research, we found that cancer has a significant economic burden, with costs accounting for hundreds of billions of dollars of US Health Care expenditures. The rising number of cancer cases and the associated healthcare costs have become a major concern for healthcare systems and policymakers. We gathered data with factors such as alcohol/ substance abuse and air pollution as potential factors of cancer - so that public policymakers and regulatory bodies could enforce some insights from data-centric recommendations. Using methods of exploratory data analysis, data cleaning, modeling, and other analysis, we intend to find the correlation between different variables like alcohol consumption, hospital admissions, gender, and total cancer rates, as well as answers to additional questions that may arise.

Data Sources

Hospital admissions: This dataset contains hospital admission per thousand population indexed by year from 1946 to 2019. We focused our efforts on extracting admissions per capita from 1950 to 2011. We noticed records from 1950 to 1970 had rates on a 5-year basis. This data is sourced from American Hospital Association, and the source also contains information on gender, bed counts, and stay information across the specified timeline.

Alcohol consumption: This dataset includes expenditure information for alcohol in the US in three categories: interest at home, away from home, and total expenditure per capita from 1935 to 2014. This data is sourced from the USDA ERDS site, which contains relevant sub-categorical information on liquor stores, food stores, hotels, etc. As Spirits are a factor impacting cancer pathology, we would like to explore correspondence.

Air Pollution: This dataset sourced from CEDS contains air pollution data by country from 1750-2019 across various variables such as level of nitrogen oxides, sulfur dioxide, carbon monoxide, organochlorines, non-methane volatile organic compounds, black carbon, and anhydrous ammonia, etc., which could relate to the environmental factors of prognosis and incidence of any health conditions. Air pollution could affect the fine particle distribution in breathable air and connect to an increased risk of acquiring cancer; hence we considered this dataset worth exploring with cancer death rates.

Cancer death rates: This dataset contains cancer death rates by type and year in the US from 1930 - 2011, sourced from the American Cancer Society. We focused on variables relating to total cancer rates by gender and cumulative rates with other factors in our final dataset across the timeline. As we have cause factors with death rates - we found this dataset would help us associate and assess delayed effects of exposure with the outcome (i.e., cancer-related death).

Population data: This dataset contains demographic information from 1950 to 2100 for each Country from UN Website. We only chose variables such as Total population and population by gender in thousands to relate and generalize the rates from other datasets. This could also inform us to see the rate of incidence of specific conditions per population and by year.

Smoking data: This dataset has smoking data related to many countries from National Statistics, including the US, from 1875 to 2015. The columns have data on average cigarette sales per adult over time. We will focus mainly on data related to cigarette purchases in the US within specified years and normalize it by population or year.

Data Manipulation Methods

The data from each of the sources are read in, and some basic cleaning of column names, selecting data from the correct country, and subsetting relevant columns are performed. Next, time lag features are engineered before we further subset the data to reflect years from 1950 to 2011. This range of years is selected as it covers the largest range of data, which overlaps all of our data.

Following this, we tackle the data missing in the range of 1950 to 1970 for the hospital admissions data, as mentioned in the previous section. Since the data in this time frame is only

recorded every 5 years, interpolation appears to be a suitable method of imputing our missing values. We visualize the missing data as well as 3 different interpolation methods – linear, quadratic, and cubic. As we see from Fig 1, the linear interpolation appears to perform decently at filling in our missing data; however, the quadratic interpolation seems to perform better visually. As for the cubic interpolation, we do not see any meaningful improvement over the quadratic interpolation, and as such, we elect to use the lower-order quadratic interpolation in order to avoid overfitting.

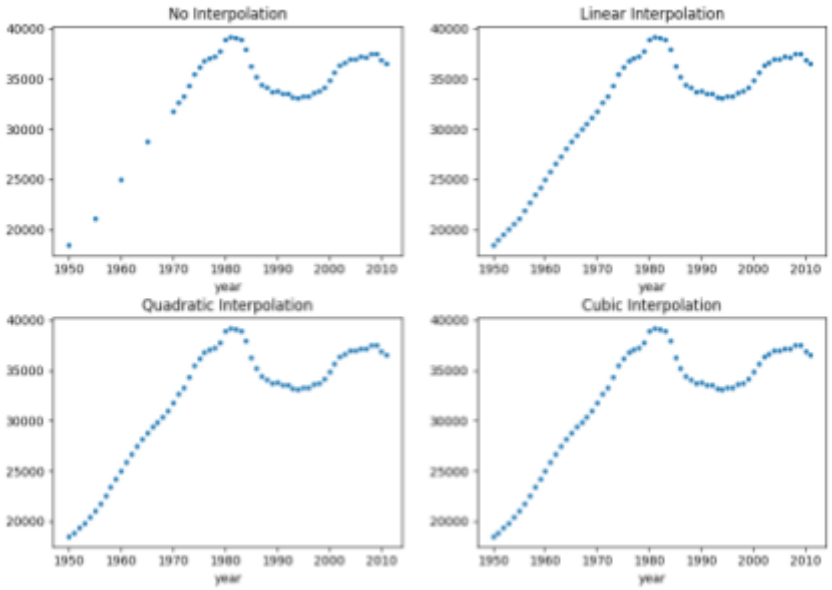


Fig 1: Plots of different interpolation techniques for hospital admissions data

After the interpolation is performed, we proceed to merge all of the separate data frames by joining via their respective years. For the hospital admissions data, we are then able to divide the total number of admissions by the population in order to obtain an admissions per capita variable, which is more useful for comparison across different years. Next, we have the cancer death rates for different cancer types split by gender in our dataset, and we are interested in the total cancer death rate. A naive approach would be to simply average the rates between the male and female rates for each cancer type, but this assumes that the split in the population between males and females is exactly 50/50. Thus, we multiply the rates for each cancer and for each gender by their corresponding gender population size to obtain the number of deaths for that cancer type and gender, sum the number of deaths, and finally divide by the total population in order to obtain a combined cancer death rate per cancer type across the entire population. We then summed these rates in order to obtain a cancer death rate for all cancer types. Lastly, each of the alcohol expenditure variables are divided by the total population so as to attain an alcohol expenditure per capita value. We note that we do not need to worry about inflation in these values as the expenditures are measured in constant 1998 US$.

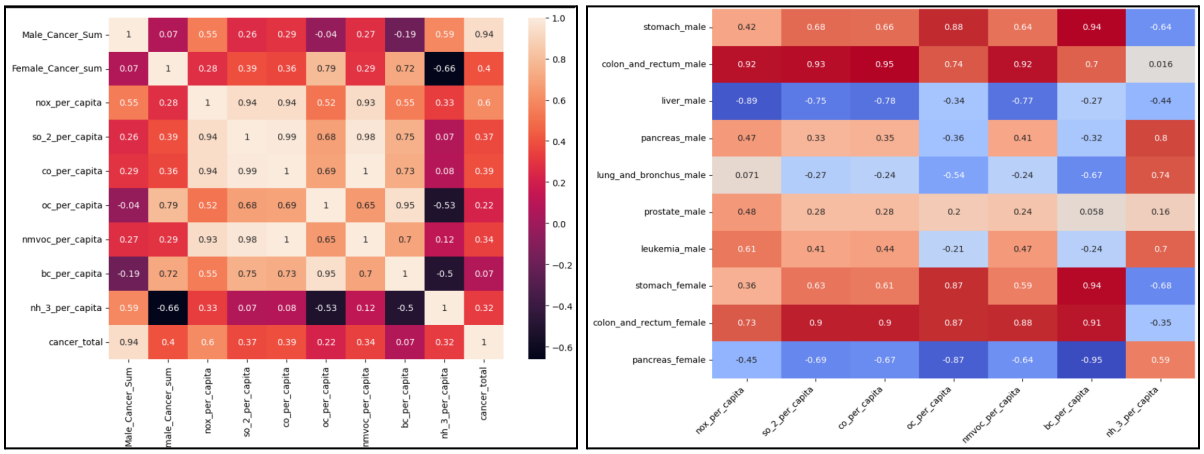
Exploratory Data Analysis

The original cancer data contained multiple columns pertaining to gender, cancer type, and total cancer death rates. To perform this part of the EDA, we split the cleaned data frames into two separate data frames based on cancer types per gender and a collection of different cancer types/ total cancer death rates for the whole population. We did this to recognize potential

correlations between gender and cancer to research further associative factors like alcohol, air pollution, and other substance abuse. We would also like to input lag features for previous year correlations as some of these factors can take years to show any associative symptom.

Gender Cancer Death Data

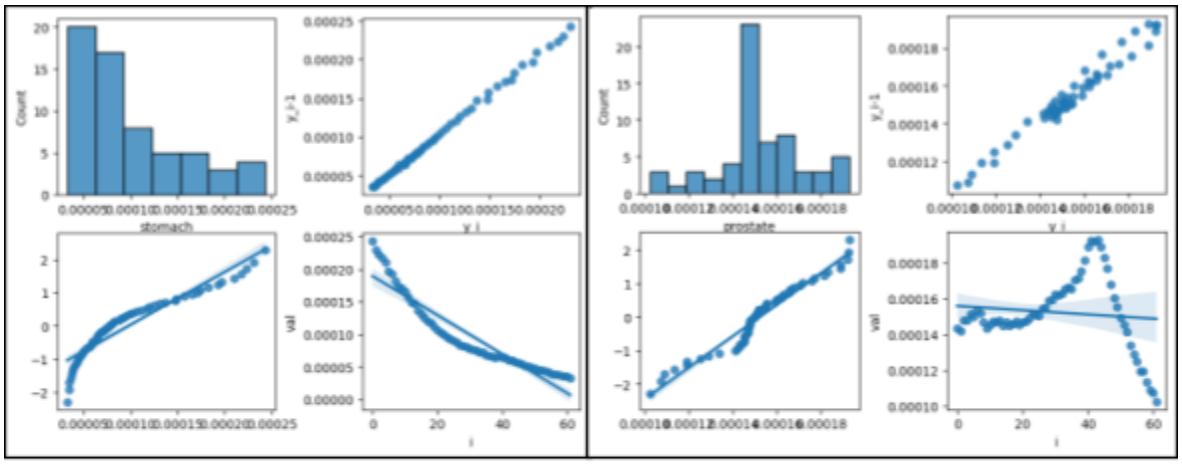
We tried to analyze cancer death rates by gender to associated factors and found that some specific cancer types are highly related by gender to factor. Some cancers, such as the breast and prostate, do not have counterparts. During our correlation analysis, we found nh3 per capita has a higher correlation with male cancer death rates than the higher correlation of female cancer death rates with organochlorines and black carbon emission per capita rates. Other factors, such as variations in exposure patterns, behaviors, or physiological traits between the sexes, would also be associated with cancer death rates. Given the limitations of correlation analysis and the need for further research to establish any causal connections or fully comprehend the patterns, it is crucial to consider additional variables and factors influencing the relationship and interpret the results cautiously.



Figs 2 & 3: Correlation heat maps of Total cancer rates by gender (left) & individual cancer rates by gender (right) vs. air pollution metrics, respectively.

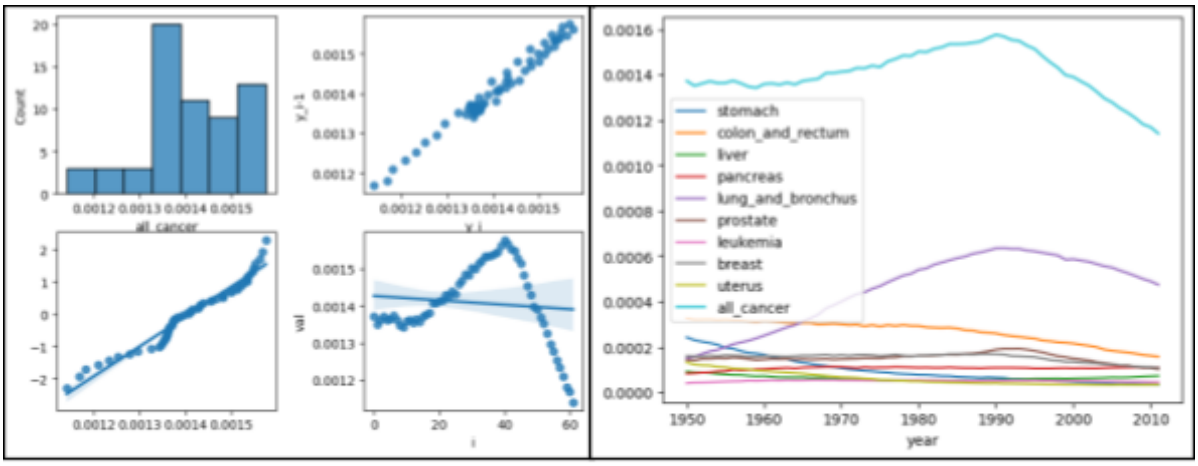
Total Cancer Death Data

We split the original cleaned data frame into a new data frame called "total_cancer," which contained total population death rates for stomach cancer, colon/ rectal cancer, liver cancer, pancreatic cancer, lung and bronchus cancer, prostate cancer, leukemia cancer, breast cancer, uterus cancer, and total cancers. We used a multiple plot function taken from our in-class notebook to showcase distribution histograms, lag plots, QQ plots, and reg plots. Most of these columns contained very abnormally distributed histograms with either slight or heavy skews. An example of this is Figs 4 & 5, with total stomach cancer and prostate cancer skewed differently.



Figs 4 & 5: Multiple plots for total right skewed stomach cancer (left) and total prostate cancer (right) showcasing distribution and yearly trends.

The separate reg plots for each cancer explains that there are strong spikes that are directly correlated to yearly data for many different types of cancers. An example of this is shown below, where total cancer death rates spiked around 1980-1990 and fell years later. To the left (Fig 6), we are showing the multiple plot and distribution of all cancer death rates. On the right (Fig 7), we are showing the overall trend of all cancer death rates. Other cancers, like lung cancer and pancreatic cancer, showed similar death rate trends, as we can see in Fig 7.



Figs 6 & 7: Multiple plots for total cancer showcasing distribution and yearly trends (left). Yearly trends for all cancers (right).

We can also verify in the visualization that some other cancers, like uterus cancer, liver cancer, leukemia, and others stayed relatively constant. However, we'd like to move forward with testing lung cancer due to the higher rate it has compared to the others in regard to total cancer rates. In all, we are assuming that lung cancer has a higher weight to total cancers than any other category. With this, we wanted to test the correlations between different types of cancers to notice any other trends there may be. Comparatively, we ran a heatmap shown in Fig 8 to observe the different correlations and were surprised at the fact that breast and prostate cancer had higher correlations to total cancer than lung cancer did. Lung cancer had a 0.42 correlation,

inferring a small correlation to total cancer. However, prostate cancer had a 0.95 correlation, while breast cancer had a 0.76 correlation, inferring stronger correlations to total cancer.

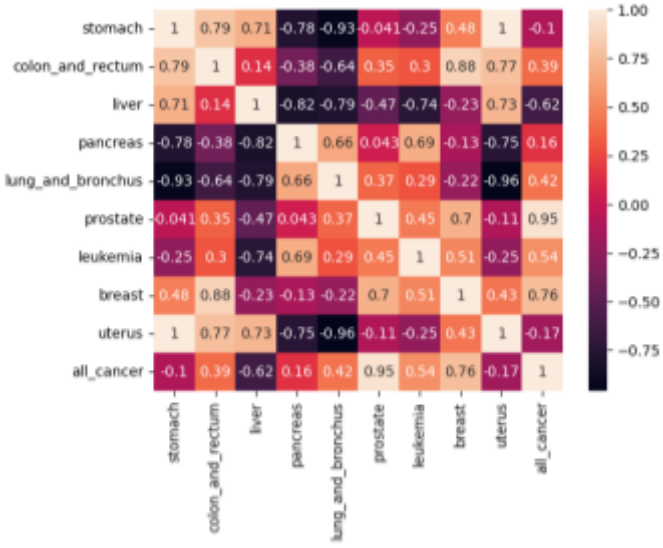


Fig 8: Correlation heat map of each total cancer type in comparison to each other

It's really important to note that correlations do not measure weight but the slope of those two variables. Referencing the reg plots we used above for both total cancer and prostate cancer, we can immediately see the similarity between the graphs and their trends. We can see other correlations like the correlation of '1' between uterus and stomach cancer and other higher (or lower) correlations between different features, although, for the purpose of our motivation, we will not be expanding on this at the time.

Overall, this EDA has led us to investigate further relationships between lung cancer and other factors like air pollution and alcohol usage and helped us navigate the addition of other data sources like average smoking per year. Hopefully, this point of analysis can assist us in better understanding the different associations between cancer and environmental factors.

Further Analysis and Modelling

As with diseases such as cancer, exposures to risk factors typically take time to manifest. Thus, we investigated the relationship between the total cancer death rates and our external factors with a lag imposed on them. Several different time frames of lag were considered, namely 10, 20, and 30 years.

First, let us consider our features which measure alcohol expenditure. We calculate the correlations between the death rate of liver cancer as well as all cancer types against the alcohol expenditure per capita features and their time-lagged versions, as shown in Fig 9. Liver cancer was singled out in this case due to the established relationship between alcohol consumption and liver disease. We see from the heatmap that there is, in fact, no significant correlation between any of the alcohol expenditure features and liver cancer death rates. However, we observe that increasing the time lags on these features shows a significant positive correlation, specifically with a lag of 30 years. We also note that we do not see these improvements occur when looking at the correlations for the total cancer death rate. This is likely due to the very different trends between the liver and total cancer death rates, as shown in Fig 7.

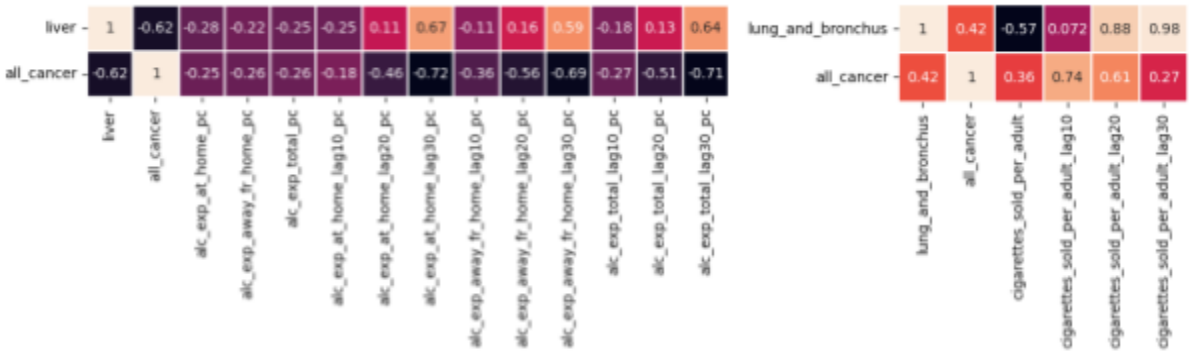


Fig 9 & 10: Correlations of Alcohol Expenditure with All Cancer and Liver Cancer Death Rates (left), and Cigarettes Sold with All Cancer and Lung & Bronchus Cancer Death Rates (right)

Next, we consider the feature measuring the number of cigarettes sold per adult per day. We calculate the correlations of this feature as well as its lag incorporated versions with the lung and bronchus cancer death rate as well as the total cancer death rate, as seen in Fig 10. We include the lung and bronchus cancer death rates due to the well-known relationship between cigarette smoking and lung cancer. We see that there's a significant positive correlation between higher lags of our feature and the lung and bronchus cancer death rate, with the highest of 0.98 being for the 30-year lag. We also see that despite this 0.98 correlation, we do not see the same improvement when looking at the correlations for the total cancer death rate.

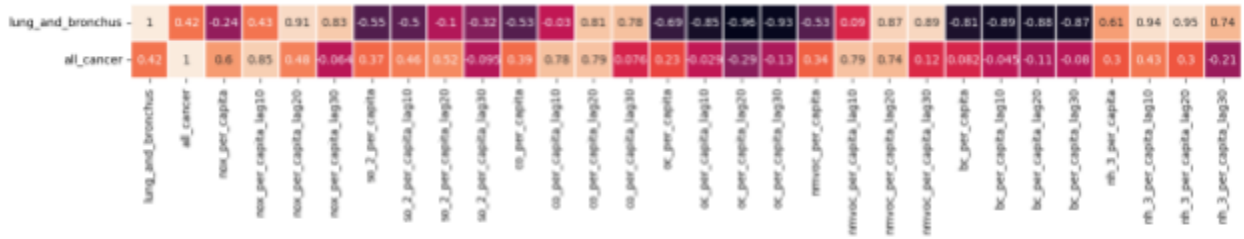


Fig 11: Correlations of Air Pollutants with All Cancer and Lung & Bronchus Cancer Death Rates

Lastly, we consider the correlations between the lung and bronchus as well as total cancer death rates with the different air pollutant features, along with their lagged variants, as in Fig 11. We include the lung and bronchus cancer death rates in this analysis as it would be the type of cancer most likely caused by air pollutants. From the heatmap, we see that Nitrogen Oxides (nox) and Carbon Monoxide (co), with a lag of 20, had the highest positive correlation with the lung and bronchus cancer death rates. As for non-methane volatile organic compounds (nmvoc), the variable with a lag of 20 years had a very slightly lower correlation with lung and bronchus cancer death rates compared to the 30-year lag variable, but the 20-year lag variable had a significantly higher correlation with the total cancer death rate. We see this also happen with the 10 and 20-year lag ammonia (nh3) variables, where the 10-year lag variable had a lower correlation than the 20 with lung and bronchus cancer death rates but a higher correlation with the total cancer death rate. As for Sulfur Dioxide (so2), organochlorines (oc), and black carbon

(bc), we do not see any meaningful correlations with lung and bronchus cancer death rates regardless of the period of lag. We do, however, see that for so2 with a lag of 20, it has a relatively high positive correlation of 0.52 between it and the total cancer rate. It is interesting to note that with all the pollutants with a significant time lag variable, the lag period was 20 years.

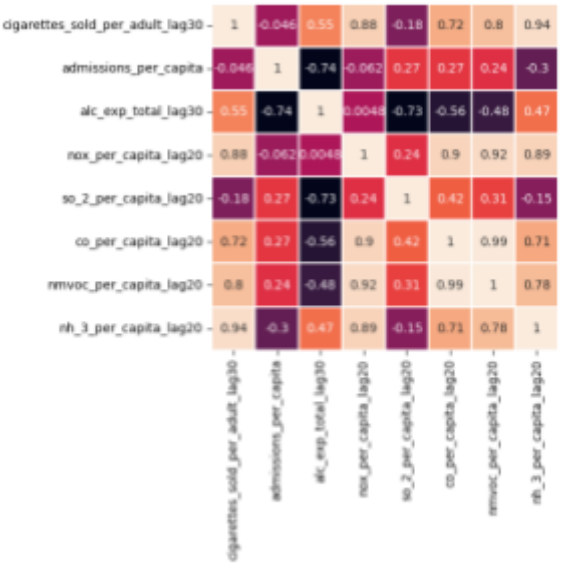


Fig 12: Correlations amongst candidate predictors

From these analyses of features with different time lags, we seek to model the total cancer death rates using a linear model with the predictors: number of hospital admissions per capita, total alcohol expenditure per capita with a lag of 30 years, cigarettes sold per adult with a lag of 30 years, and the 20 years lagged per capita concentrations of nox, so2, co, nmvoc, and nh3. Before modeling, the correlations between all of our predictors are calculated, as shown in Fig 12, and we elect to remove the features for nh3, co, nox, and nmvoc due to their high correlations with multiple features, which would introduce high levels of multicollinearity in our model if included. We choose to employ a linear model due to its lightweight and easily interpretable nature.

Additionally, the variable reflecting alcohol expenditure only has data from 1935, and introducing a time lag of 30 years into the variable causes the first 15 years of our dataset to contain missing values, and we are left with 47 observations; thus, a linear model is suitable, as more complex models are likely to overfit and have poor generalizability.

OLS Regression Results

Dep. Variable:	all_cancer	R-squared:	0.946
Model:	OLS	Adj. R-squared:	0.941
Method:	Least Squares	F-statistic:	182.9
Date:	Sun, 09 Apr 2023	Prob (F-statistic):	5.61e-26
Time:	18:37:03	Log-Likelihood:	428.81
No. Observations:	47	AIC:	-847.6
Df Residuals:	42	BIC:	-838.4
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|
Intercept	0.0006	0.000	5.728	0.000
cigarettes_sold_per_adult_lag30	5.618e-05	3.43e-06	16.391	0.000
admissions_per_capita	1.399e-06	3.69e-07	3.793	0.000
alc_exp_total_lag30_pc	-2.461e-09	3.08e-10	-7.989	0.000
so_2_per_capita_lag20	2.002e-06	4.27e-07	4.688	0.000

Fig 13: OLS Regression Summary (Confidence Intervals & Diagnostics Truncated)

Thus, a linear model for the total cancer death rate is fitted with the following predictors: the number of hospital admissions per capita, cigarettes sold per adult with a lag of 30 years, total alcohol expenditure with a lag of 30 years, and so2 concentrations per capita with a lag of 20 years. The summary of the model is seen in Fig 13 and we see that the p-values for each of the variables are extremely low, being less than 0.001, indicating that all of our variables are significant in predicting the total cancer death rate. We also see that we have a very high adjusted R-squared value of 0.941, indicating that our model is able to capture 94.1% of the variability in the data. It is important to note that these findings are associative, and no causal relationship can be drawn between our response and predictor variables.

Conclusion

In conclusion, our analysis focused on finding correlations between cancer death rates and etiogenic factors across timelines focusing on the US population based on available data and our research. The key findings emphasize that carcinogenic factors which relate to carcinogenesis and the progression of cancer rates don't have an immediate result of death but do relate with it to a time gap which makes it crucial to palliate and reduce the risk to decrease the possibility of unfavorable outcomes.

Throughout our process, we have performed many different merges of the datasets, analyzation of different data, and exploratory factor analysis to assist us in furthering our decisions. After our exploratory analysis, further analysis was performed in order to investigate the effects of incorporating a time lag into our features, showing that there is significant value to such a form of feature engineering when modeling a disease. Lastly, an explainable model was formed, in which we see all of our variables are statistically significant.

With modern medicine diving into more customized care, putting the patient in the center of the care delivery utilizing clinical data to make informed decisions, we could potentially help hospitals and public health services predict the magnitude and severity of different diseases in the future, given historical data. This will allow the health sector to better plan and prepare itself to provide patients

Finally, while interpreting this analysis to make policies reflecting the general population, it’s important to consider that the data we have collected, used, and analyzed might not be representative of the broader population. As we have utilized publicly sourced data, we acknowledge the limitations and biases that this analysis has on interpretation. We do, however, hope that our research is useful in furthering research and affecting global health overall.

References (APA Format)

1. What Is Cancer? (2021, October 11). National Cancer Institute. https://www.cancer.gov/about-cancer/understanding/what-is-cancer#:~:text=Cancer%20is%20caused%20by%20certain,tightly%20packed%20DNA%20called%20chromosomes.&text=Cancer%20is%20a%20genetic%20disease,how%20they%20grow%20and%20divide
2. Financial Burden of Cancer Care. (n.d.). Cancer Trends Progress Report. https://progressreport.cancer.gov/after/economic_burden
3. What Causes Cancer? | American Cancer Society. (n.d.). https://www.cancer.org/healthy/cancer-causes.html
4. Alcohol Use and Cancer. (n.d.). https://www.cancer.org/healthy/cancer-causes/diet-physical-activity/alcohol-use-and-cancer.html#:~:text=Liver%20cancer%3A%20Long%2Dterm%20alcohol,the%20risk%20of%20liver%20cancer.
5. What Are the Risk Factors for Lung Cancer? | CDC. (n.d.). https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm#:~:text=People%20who%20smoke%20cigarettes%20are,the%20risk%20of%20lung%20cancer.
6. Nature Publishing Group. (2023). How air pollution causes lung cancer — without harming DNA. Nature. https://doi.org/10.1038/d41586-023-00989-z
7. Links to Data
    a. https://www-statista-com.proxy.lib.umich.edu/statistics/459718/total-hospital-admission-number-in-the-us/?locale=en
    b. https://github.com/owid/owid-datasets/tree/master/datasets/Alcohol%20expenditure%20in%20the%20USA%20long-term%20(USDA%2C%202018)
    c. https://github.com/owid/owid-datasets/tree/master/datasets/Air%20pollution%20emissions%20(CEDS%2C%202022)
    d. https://github.com/owid/owid-datasets/blob/master/datasets/Cancer%20death%20rates%20in%20the%20US%20over%20the%20long-term%20-%20American%20Cancer%20Society/Cancer%20death%20rates%20in%20the%20US%20over%20the%20long-term%20-%20American%20Cancer%20Society.csv
    e. https://population.un.org/wpp/Download/Standard/MostUsed/
    f. https://ourworldindata.org/grapher/sales-of-cigarettes-per-adult-per-day

Statement of Work

We used GitHub to share the code and divide parts before compiling the final source code. We used Shared Drive for reports and background research work to share our thoughts.

Who did what:

- Amit Das: Data extraction, cleaning, merging data sets, exploratory data analysis, literature review
- Pratap Gude: Exploratory data analysis, data merging and cleaning, making sure that coding is in PEP-8 guidelines, modeling, and model evaluation
- Eman Wong: Modeling and model evaluation, statistical analysis, Project report writing, literature review

Assessment of Collaboration:

Overall the project timeline suited our timeframes, we were able to start the project early on, gather data and understand the strengths and weaknesses of the final data frame (for instance normalizing rates across variables, dealing with missingness). Apart from the collaborative platforms, we were able to meet frequently to update the project status with each other. We believe we divided the work to the best of our strengths and were able to learn from each other during the process.

Areas of Improvement:

We thought we could have gathered some feedback midway through the project from the teaching team to recognize any lagging areas. Overall, we were able to meet the faculty individually and ask some questions but not as a group. We could have also correlated more time to the project to get better results, but due to the limitations of our coursework, this was difficult to maximize.