Cancer Prevalence and Risk Factors Associated

by:
Matt Chang, Blake Maxwell, Jacob Peterson

Summary of Research Questions:

1) **How has cancer prevalence changed over time for certain populations?**
   We will look at sex, and age related data and their trends of prevalence for groups of cancer types over a 20 year time window. We can then assess patterns in different populations over time by isolating a population and graphing the result.

   Answer: We see a general increase in cancer prevalence from 2000-2018, and when looking at people under the age of 50, a relatively large increase in cancer diagnoses over the time period while there is a negligible difference in increase between sexes.
2) **Does alcohol consumption correlate with cancer frequencies?**

   Answer: Our research found no correlation between varying levels of alcohol consumption and cancer frequencies. However, these results were likely a result of the limitations of our data, and does not disprove alcohol's well-established role as a carcinogen.
3) **Do smoking behaviors (nicotine and tobacco use) correlate with cancer frequencies?**

   Answer: Our research has found correlation with a number of risk factors and tobacco-related cancers. The most significant of these were ever smoking cigarettes or ever smoking pipe or hookah. Regularly smoking cigarettes showed the highest frequency of lung cancer out of the regular tobacco users.
4) **From the determination of risk factors and their weight on cancer diagnosis, how do we account for the age groups that are largely excluded for the risk groups that contribute toward major cancer types?**

   Answer: Sex is a factor in being diagnosed with liver or lung cancer and young people still are a considerably large proportion of cancers in the U.S. While breast and skin cancer are most prevalent, lung and liver cancers associated with certain risk factors are prevalent as well.

Motivation:

Everyone has been or known of someone who has dealt with the seriousness of cancer. Studying cancer prevalence and looking for unsettling trends is critical in order to drive further research in treatments and potentially a cure for cancers. We grow up hearing about risky behaviors that may affect our health in the future but analyzing data on these behavioral risk factors allows more educated decision making. According to the FDA, "Though only a small percentage of teens used e-cigarettes in 2011, 28 percent of high school students and 11 percent of middle school students used e-cigarettes by 2019." E-cigarette and cigarette use and its effects on the lungs or other organs is still an active area of research in the industry. An example like this and the countless others that are known carcinogens to humans, though brought up in schools and media outlets, may not capture the full scope. Through this project we want to analyze cancer prevalence trends and investigate risk factors that may play a role in someone being diagnosed with cancer. Making this data digestible and educating community members as it pertains to cancer prevalence and risk factors is a fundamental goal of this project.

As college students, we are in one of the largest groups of alcohol/ binge drinking as well as smoking/vaping use. There is value in looking at trends and numbers across age groups to understand where trends are changing and how this may lead to more or less cancer in the future. Smoking can predispose a person to lung cancer while extended alcohol use is seen to increase risk of liver cancer. Breaking these subgroups down allows people to understand their risks before they partake in these activities. Laws can be structured around evidence based research and medical professionals can use data analysis to better inform their practices. Cancer can be devastating for our society and until there is a cure, research into the aforementioned is imperative.

Datasets:

- CDC United States and Puerto Rico Cancer Statistics, 1999-2018, 0-50 Years Old, both sexes, All cancer type, grouped by cancer sites, year, age groups, and sex.
  https://wonder.cdc.gov/controller/datarequest/D172;jsessionid=92B397CFCD1A20E618401031A3BC
- CDC Single-Race Population Estimates 2010-2020 and single-year age request
  - Broken down into unisex, male, and female datasets for analysis

  https://www.census.gov/data/tables/time-series/demo/popest/intercensal-2000-2010-national.html

- NHIS 2018 Sample Adult Survey
  - Consists of a questionnaire given to adults with questions regarding health metrics, demographic attributes, and risk factors.

  https://www.cdc.gov/nchs/nhis/nhis_2018_data_release.htm
  Found under Data files → Sample Adult File → CSV data

Methodology:

- **Data Manipulation**
  The population data CSVs were downloaded from the CDC national census data for the year ranges 2000-2010 and 2010-2019. They were combined manually to be a single file, and then split into a CSV for each the female, male, and unisex population data. The

sample study CSV was downloaded directly from the NHIS link described above. The cancer data was requested from the CDC site described above. The cancer and test data was downloaded in txt format, loaded into Microsoft Excel to transform into CSV format for use in code.

- **First Research Question**

  To prepare for looking at cancer prevalence, we need to clean up and convert the datasets to usable forms. Once the files are in proper csv format, we begin by adjusting the age group labels of the 3 datasets regarding population estimates to match the CDC Cancer Statistics dataset. The population estimates are then used in conjunction with count data by cancer to create frequencies of cancer by the population numbers within the United States. We append this metric as a new column in the Cancer Statistics dataset. These frequencies take into sex, age group and cancer type as subgroups. For instance a male, aged 21 with liver cancer will be divided by the total estimated population of males, aged 20-24 in the United States. By changing count data into frequency data, we will be able to see a more accurate representation of the breakdown of cancer in the United States. Using frequencies, we remove years outside of 2000-2018 for consistency in both datasets. Finally, using Plotly we can graph the frequencies of cancer over this given time period and look at variability.

- **Second Research Question (Alcohol)**

  The first risk factor we explored was alcohol use. We chose to explore alcohol consumption due to its established role as a carcinogen. According to a study conducted by researchers at the National Cancer Institute, 3.5% of cancer deaths in the United States in 2009 were alcohol related (Nelson, 2013). The 2018 NHIS Adult Sample Data asked participants if they had ever been diagnosed with any of 30 different cancers. So, we began by first filtering the data to include only seven of the original thirty different kinds of cancers. The seven selected cancers were of types that have been associated with alcohol consumption according to the National Cancer Institute. Another filter we applied to the NHIS dataset was to include the ‘ALCSTAT’ column, which categorized participants into groups based on their amount of alcohol drinking consumption. We also decided to only include data from participants over the age of 60, due to the fact that most cancers we were looking at usually first appear in individuals over the age of 60. After the data had been filtered, we subsequently calculated the frequencies of each type of cancer within each ‘ALCSTAT’ group, along with the combined frequency of each alcohol-associated cancer for each ‘ALCSTAT’ group. Using this data, we created two visualizations using the Plotly library. The first was a bar chart comparing the combined frequency of all alcohol-associated cancers between each ‘ALCSTAT’ group. The second visual included 7 sub-plotted bar charts-one for each ‘ALCSTAT’ group. Each subplot compared the frequencies of each type of alcohol-associated cancer within a given ‘ALCSTAT’ group.

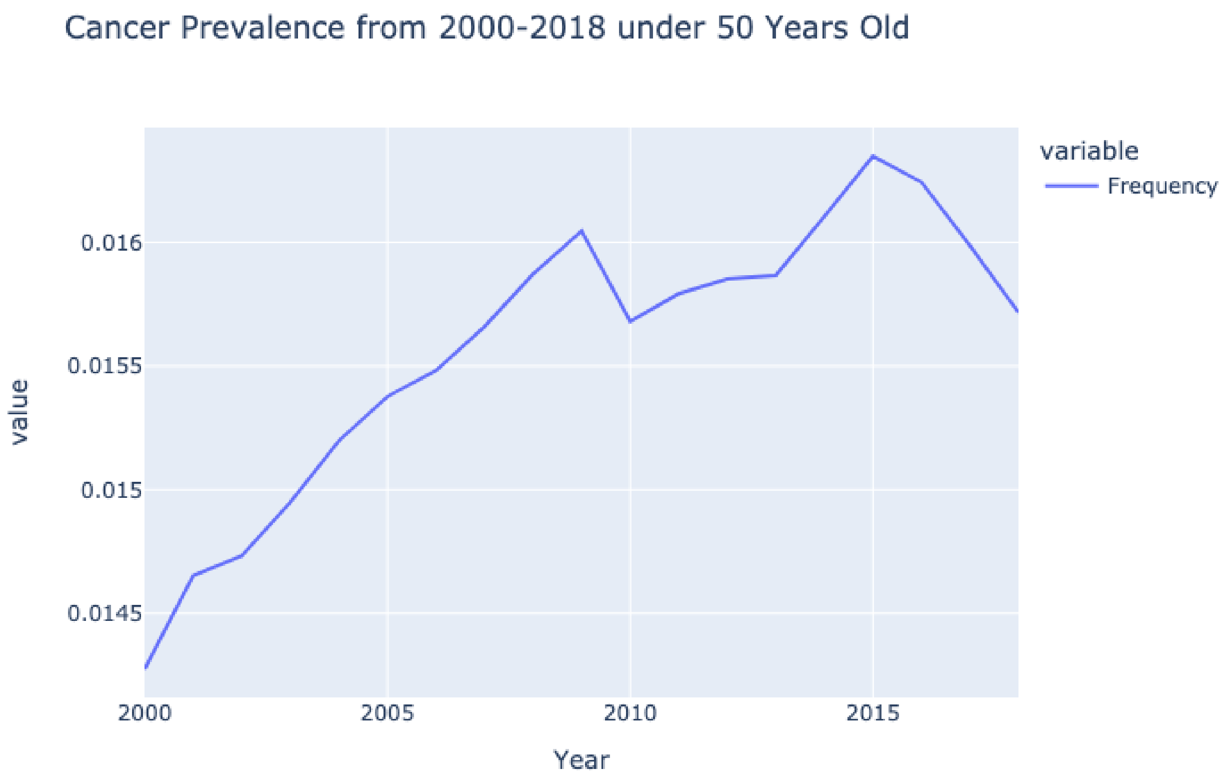
- **Third Research Question (Smoking)**

The second risk factor we explored was nicotine use. We chose this because of its significant correlation with respiratory and other cancers. We then selected only the columns relating to cancers from tobacco use. The columns had to be manipulated to be analyzed including establishing variables that can be used to plot. A bar graph will be the best visualization as it will show the frequencies broken down by tobacco users and the regularity of their usage. This question will mirror the methodology set forth by the second research question about alcohol.

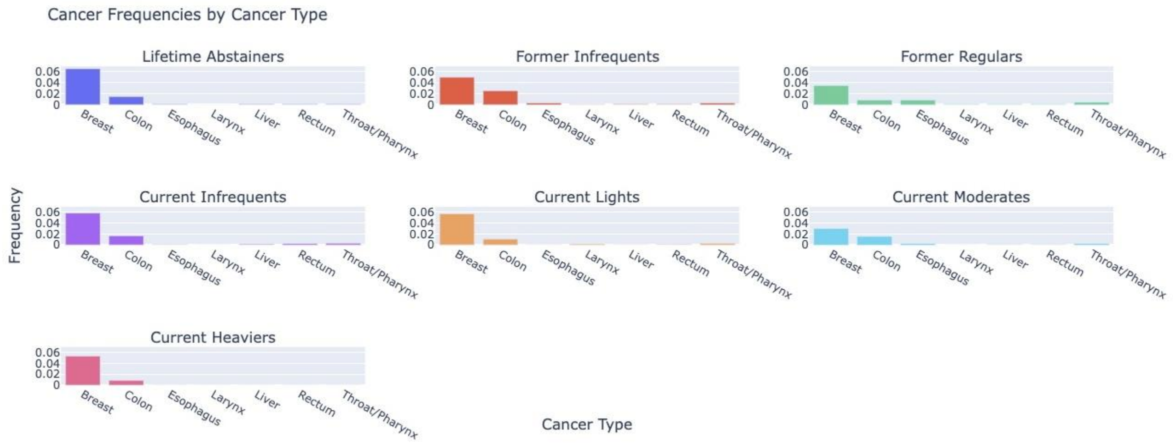
- **Fourth Research Question**
  After looking at the results from both the third and fourth research question, we want to zoom out our focus of risk factors and look at large contributors to these results. With the narrowed down 2018 NHIS Adult Sample Data, we graph the incidence of cancer by cancer type for 2018. Part of the reason we chose to use 2018 data is for comparison to the cancer prevalence over time and this year being the most recent possible public data. It allows us to give the most relevant outlook of cancer in the United States especially for those young people who may not be included in as many cancer studies. With robust metrics on alcohol and tobacco use, we will break down the most likely cancer that can be caused by these drugs. For both lung and liver cancer, we create subset datasets of just numbers associated with these two cancers. Removing age groups with lack of data for certain years, or for age groups with insignificant data is the next step. For both lung and liver cancers, we will create individual interactive visualizations with Plotly. Using this powerful library gives more in depth information about the breakdown of these cancers. Each visualization will contain facet plots split by age group. These will be frequencies of cancer incidence over time per age group with different area plots by sex.

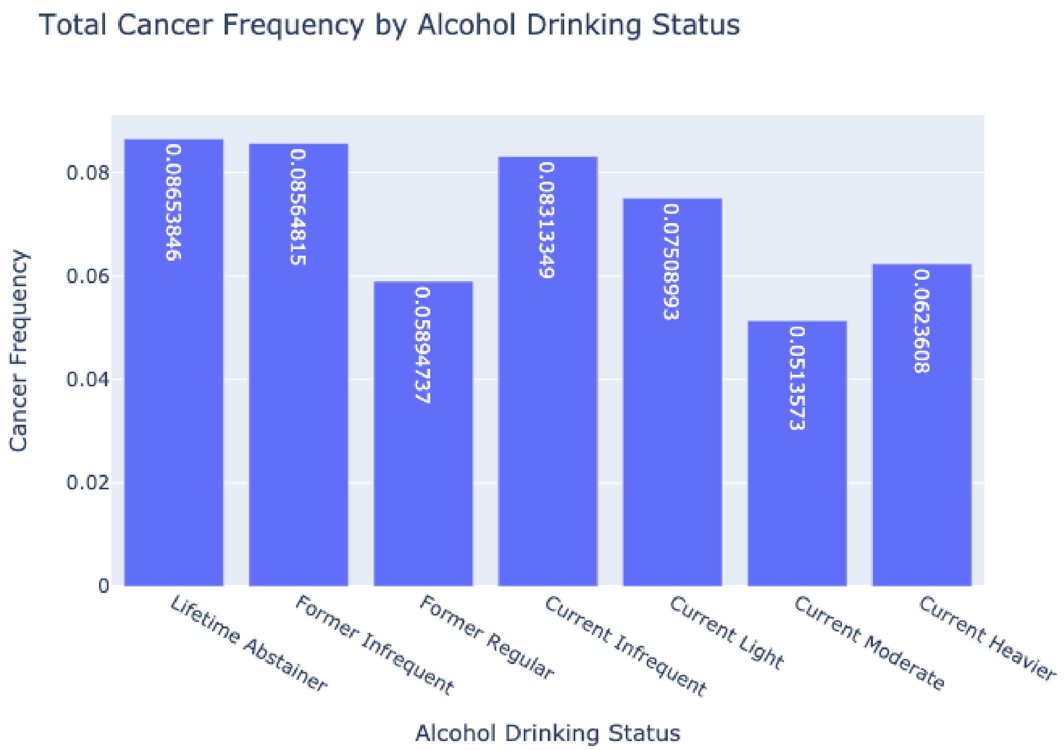
Results:

- **First Research Question**
  Cancer prevalence over time for people aged 50 or under shows steady increase from 2000-2008 before dipping down for a 2 year time window. The overall highest frequency over these years is in 2015 where 1.63% of people in the United States were diagnosed with or had active cancer in this year. Cancer incidence increased 14.78% over the first decade and a half of the 21st century for people aged 50 and under. This is quite astonishing considering the small time frame of data analysis completed. It is likely that cancer incidence will ebb and flow like seen in the line graph but remain at a general increase over larger spans of time. Seeing that cancer prevalence is increasing in younger individuals is concerning as cancer can be portrayed as age related complications. This first research question is the gateway into looking at specific cancers and risk factors that may increase this incidence in our remaining research questions.

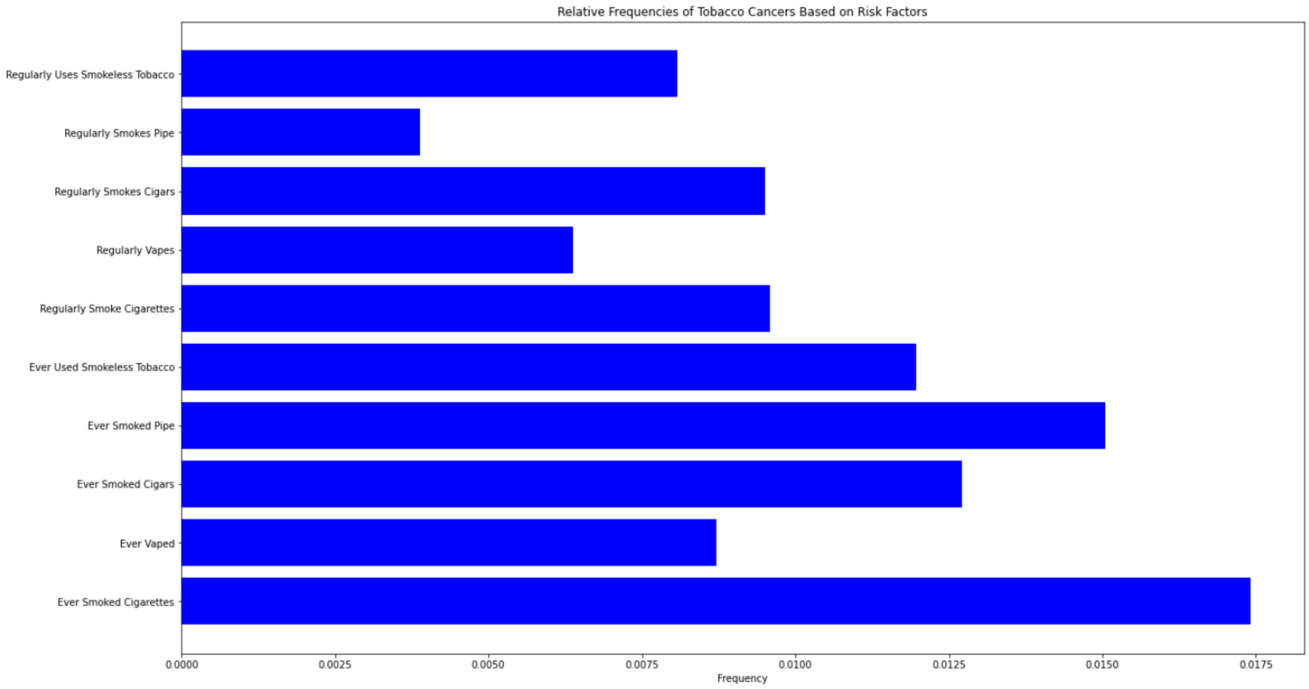


- **Second Research Question (Alcohol)**
  The first bar chart compared the combined frequencies of all alcohol-associated cancer types between ‘ALCSTAT’ groups. This visual showed no correlation between different levels of alcohol consumption and cancer rates. The second visual included seven sub-plotted bar charts comparing the frequencies of each type of alcohol-associated cancer within each ‘ALCSTAT’ group. This visual revealed no significant differences in rate of each cancer type between different ‘ALCSTAT’ groups.



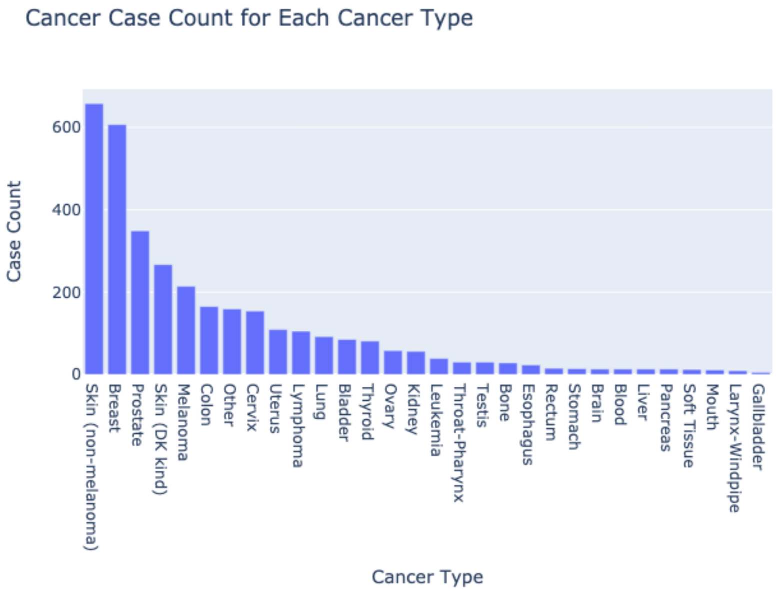


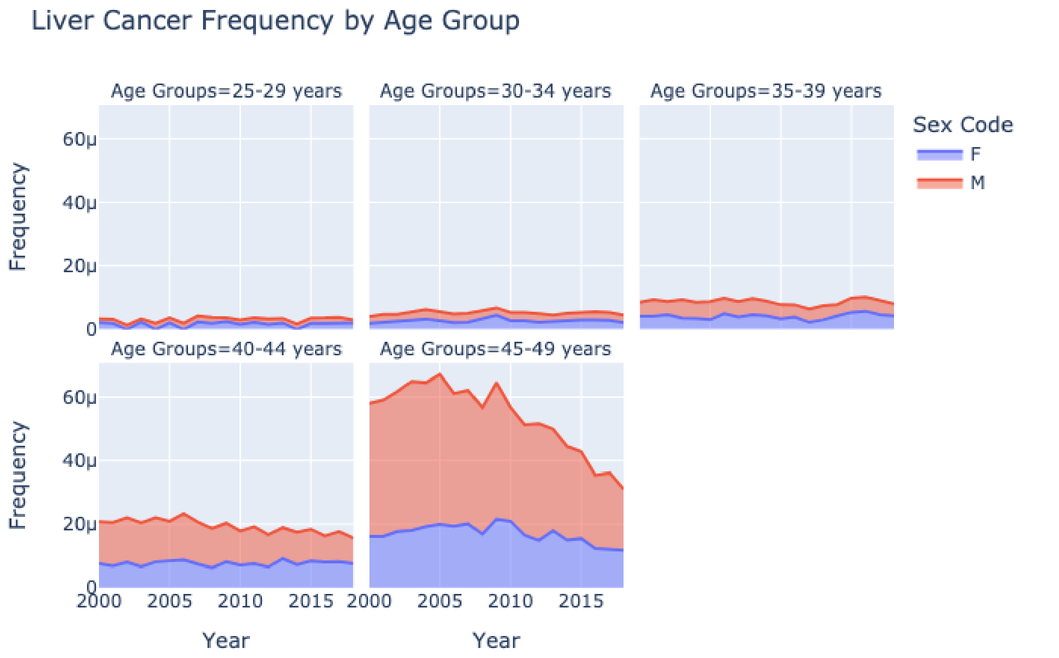
- **Third Research Question (Smoking)**
  The bars in the graph represent the number of people who have the given risk factor as well as respiratory cancer divided by the total number of people with the risk factor. This is the frequency metric on the x-axis. Almost all of the "ever smoked" etc. variables at the bottom of the bar chart show higher frequencies of lung cancer which is a bit different than expected. We assumed there would be more frequency of cancer in the regular use of tobacco groups but this is not the case. Cigarette use lead all of the other types of tobacco in terms of ever smoking or regular smoking of cigarettes. We believe that seeing the bars at the bottom of the chart being larger is due to the number of people in these groups in comparison and with lack of very large surveying, these numbers can be easily skewed.

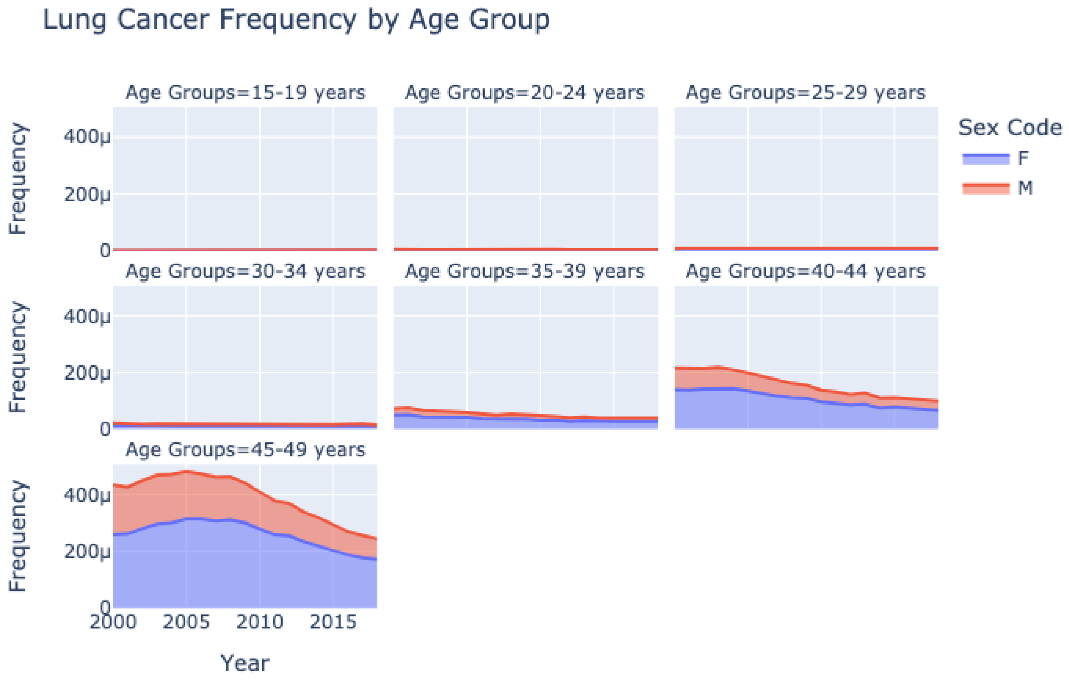


- **Fourth Research Question**

  In terms of the cancer case count by cancer type in the year 2018, skin and breast cancer are by far the most prevalent in the United States. While these are not commonly mentioned cancers with respect to alcohol and drug use, it is interesting to see the magnitude of these cancers in comparison to others. Lung cancer is many times more common than liver cancer but liver cancer does still make up the 25th most common cancer in the US, while lung cancer is 11th according to our data. Breast cancer is 47x higher in incidence than liver cancer. Accounting for more age groups including those over the age of 50 shows which cancers have the most impact on people living in this country. On the other hand, both lung cancer and liver cancer are seen in greater proportions later in life. The younger age groups with lung and liver cancer actually show stark increases in incidence after the age of 40. These cancers surprisingly contradict the overall increase of cancer incidence in the U.S. as both liver and lung cancer show declining frequency in people aged 40-49. The facet plots show stacked layers of frequency by sex which allows for the combined frequency to be seen in the same plot. Males make up 70.1% of liver cancer frequency in people 45-49, however, females make up 65.4% of lung cancer frequency in people 45-49 at their respective highest points. A clear division by sex is intriguing because sex is determinably a factor in risk of being diagnosed with either of these cancers.


Cancer Case Count for Each Cancer Type
Case Count
600
400
200
0
Skin (non-melanoma)
Breast
Prostate
Skin (DK kind)
Melanoma
Colon
Other
Cervix
Uterus
Lymphoma
Lung
Bladder
Thyroid
Ovary
Kidney
Leukemia
Throat-Pharynx
Testis
Bone
Esophagus
Rectum
Stomach
Brain
Blood
Liver
Pancreas
Soft Tissue
Mouth
Larynx-Windpipe
Gallbladder
Cancer Type


Liver Cancer Frequency by Age Group
Age Groups=25-29 years
Age Groups=30-34 years
Age Groups=35-39 years
Age Groups=40-44 years
Age Groups=45-49 years
Frequency
60μ
40μ
20μ
0
2000
2005
2010
2015
Year
Sex Code
F
M


Lung Cancer Frequency by Age Group
Age Groups=15-19 years
Age Groups=20-24 years
Age Groups=25-29 years
Age Groups=30-34 years
Age Groups=35-39 years
Age Groups=40-44 years
Age Groups=45-49 years
Frequency
400μ
200μ
0
2000
2005
2010
2015
Year
Sex Code
F
M

Impact and Limitations:

- **First Research Question**

  Research into cancer prevalence has the potential to impact people around the U.S. and the globe. Assessing trends of cancer diagnosis, and especially those that slow alarming statistics, increases the likelihood of the government writing laws to help or provide assistance to further research and treatment methods. This report and the data analysis contained within this question cannot however be extended to all populations or what will happen in future years. The data used from the CDC is static and only looks at a limited amount of data. It will fail to account for undiagnosed cancers in the U.S. and may overlook those who do not have quality access to medical care. These visualizations purposely exclude people over the age of 50 for the purpose of our analysis and thus are not representative of the entire U.S. population. Biases exist around which providers update their statistics on cancer timely and the data cannot take into account extraordinary factors in certain years that may influence accurate reporting of prevalence. Our report aims to provide the best snapshot of frequency of cases in respect to population but more research is necessary to back up our results.
- **Second Research Question (Alcohol)**

  Considering alcohol has been previously well-established as a carcinogen, the lack of correlation we found between alcohol consumption and cancer rates is likely a consequence of the limitations of our data. For one, there were too few data points of participants diagnosed with the types of cancers that we were observing. Secondly, many of the cancers that we were looking at have many other factors other than alcohol consumption that put individuals at an increased risk of getting diagnosed with them. Lastly, a disproportionate amount of the data points we were looking at represented individuals diagnosed with breast cancer, which has many other risk factors other than alcohol consumption. For context, 607 data points represented individuals that were diagnosed with breast cancer, while only 13 data points represented individuals with liver cancer. Thus, the lack of correlation found in our research likely does not disprove alcohol's role as a carcinogen.
- **Third Research Question (Smoking)**

  Along with alcohol as a carcinogen, smoking is another carcinogen that deserves to be investigated. For us to reasonably analyze the results above, we need to take into account the limitations of a study like the one we received data from. First of all it is a snapshot of a signal year and thus we are unsure if 2018 is an anomaly or if it fits the pattern of other years without doing more research into it. For this reason, this data should be taken only as a window in time to how smoking use and lung cancer correlate. People who smoke may not keep track of how many cigarettes or tobacco they consume and thus give an unequal distribution of the actual usage of tobacco in the United States with respect to likelihood of being diagnosed with lung cancer.
- **Fourth Research Question**

From the results seen in the second and third research questions, we aim to limit the exclusion of certain populations and provide somewhat of a wider scope. From the 2018 NHIS Sample Data, looking at all cancers broken down by counts provides what our other parts do not. It highlights the most common cancers in the U.S. in the most recent year for which there is complete data. This is not to say that these most frequent cancers are the most important, but it emphasizes the impact of cancer in the U.S. yearly. In comparison to the graph about cancer prevalence over time, this new graph broken down by cancer allows a fuller picture of what cancers go into these numbers. These are independent studies however, so people must be careful drawing concrete conclusions from comparing. The lung cancer and liver cancer by age and sex visualizations are an accurate picture of frequency over time but there are limitations in this data. Like the risk factor visualizations, it does not capture people who may untruthfully report their use of drugs due to stigmas surrounding them or people who are less likely to report back to surveys because of language barriers, socioeconomic limitation, lack of access etc…

Challenge Goals:

We incorporated **multiple datasets** into our analysis for this project. One consists of counts of cancer prevalence structured around sex, race and age. Another one contains metrics of behavioral risk factors to which we will select the most appropriate factors. The last one is population data that was used for frequency measurements to replace the count data in the Cancer Statistics dataset. Comparing results from these datasets allowed us to look deeper into the potential reasons why certain cancers are higher in prevalence or why a trend is occurring. These datasets gave the option of both limited and wide scope analysis that we implemented into our code. The visualizations were also used to compare across datasets for our results. Another challenge was that our data was **not presented cleanly** in a CSV format and is instead in a txt file that will need to be converted. The data contained within each dataset was messy and not in the adequate form for analysis. We spent a great amount of time eliminating extraneous variables, changing and merging rows before the data was usable. This increased the emphasis on this challenge goal and took research into performing this data manipulation before we could create visualizations. Our final challenge goal created impressive visualizations with Plotly by learning this **new library.** More advanced visualizations provide a richer insight into the data we analyze. The visualizations in this report are unfortunately static but with our code you can see the individual levels and data points by hovering over the graphs. With how complex cancer and the associated risk factors are expected to be, a more advanced visualization will allow our results to be sophisticated.

Work Plan Evaluation:

1. Loading in data, cleaning and converting to CSV: (5 hours)

   We will need to take the txt file and convert it to a csv. In addition we will create pandas DataFrames of the data and remove all of the unnecessary variables.

2. Data Manipulation: (7 hours)

   We will need to manipulate the data into a frequency dataset by combining population data and the count data for ages, sexes and races in the cancer prevalence dataset.

3. Plotting: (4 hours)

   Plotting cancer incidence over time for certain populations. We will also use plotly to compare trends of risk factors that may lead to cancer (scatter and interactive distribution plots seem most fitting initially).

4. Report Writing and Final Checks (6 hours)

   Compiling results and making the report. This will include preparing our code for submission and meeting all requirements.

UPDATED:

For the most part our work plan evaluation was accurate. Loading in data took a bit longer than expected at around 8 hours as well as data manipulation which took much longer than expected at around 14 hours. Both plotting and report writing were spot on. We think we had a decent grasp on how long different parts of the report would take especially with the latter part. For the second part, much more research into how to manipulate the data with frequencies increased the time for this section. Combining the data and using both count of cancers and population to create these frequencies turned out to be challenging. We had not scoured the entire datasets prior to writing the proposal and did not see the full extent of the messiness of the data. Spending time looking into how to code for plotly and converting files into csv took considerable amounts of time as well.

Testing: We tested throughout the code as it developed. To be sure our results were accurate, we created a test set that was a subset of the dataset. We added a test function to make sure our functions are returning proper dataframes. Our results lie in visualizations as well as analysis from these visualizations so since we did not have any numerical output, we never needed to use the assert equals function. Checking to make sure that the test data had the same format, same behavior and comparable results allowed us to feel certain that what we are reporting is accurate. The data we used all came from government agencies with strict protocols on data collection and survey management. This ensures that the data is not falsified, altered or manipulated in the results we gain from these datasets. When problems with calculating proper frequencies arose, we were able to check that these counts are being divided properly with the correct age group, sex and year with the corresponding population data. Outliers or missing data that appeared in visualizations were removed or changed to increase accuracy.

Collaboration: The only members contributing to this report are those listed and only the course staff was consulted.

Works Cited:

Nelson, David E et al. “Alcohol-attributable cancer deaths and years of potential life lost in the United States.” *American journal of public health* vol. 103,4 (2013): 641-8. doi:10.2105/AJPH.2012.301199

“Alcohol and Cancer Risk Fact Sheet.” *National Cancer Institute*, https://www.cancer.gov/about-cancer/causes-prevention/risk/alcohol/alcohol-fact-sheet#r8.

Products, C. for T. (n.d.). *Get the latest facts on Teen Tobacco Use*. U.S. Food and Drug Administration. Retrieved March 15, 2022, from https://www.fda.gov/tobacco-products/youth-and-tobacco/get-latest-facts-teen-tobacco-use#:~:text=Though%20only%20a%20small%20percentage,schoolers%20currently%20using%20e%2Dcigarettes.