Decoding Cancer: Analyzing Cancer Factors by Country

Michelle Ly · Anne Pilling

I. Introduction

Contributions:

Michelle Ly - Visualizations, Writing, Code, Analysis, Github

Anne Pilling - Visualizations, Writing, Code, Analysis/Interpretations

Github: https://github.com/m1chelleL/STA_141B_Final

Introduction:

As a healthy and active 21 year old, I (Anne Pilling) never expected to hear the news of a cancer diagnosis. Undergoing chemotherapy throughout the summer of 2023 and into my senior year of college, questions of why this happened to me, if there could have been anything I could have done to prevent it, and how does access to healthcare affect those who aren't as lucky were rattling around in my brain. Using this as a research opportunity, Michelle and I aim to analyze and compare the data available on cancer in general to provide a sense of where we, as the United States, place among other countries.

Background:

Cancer is a global health issue affecting millions of people worldwide of all ages and regions. With over 200 types, each with their own unique characteristics and subcategories, cancer can generally be explained as when abnormal cells within the body grow uncontrollably, spreading and invading nearby tissues that can lead to the formation of tumors. Being a genetic disease, cancer is caused by changes to genes in our DNA that control the way cells grow and divide. These changes can be inherited or caused by harmful environment factors such as tobacco smoke, chemicals, radiation (like UV) but they can also happen because of errors that happen as cells divide. The body's ability to eliminate damaged cells before they become cancerous dwindles as we age leading to higher risks of cancer later in life.

Our objective is to examine how age groups affected by cancer vary across different countries, how access to healthcare influences prognosis, how physical activity levels correlate with cancer rates globally, and the impact of environmental factors like pollution. We hope our findings will contribute to a deeper understanding of cancer and its many contributing factors.

Methods & Libraries:

- ★ Web scraping
- ★ Requests

- ★ BeautifulSoup
- ★ Pandas
- ★ Pyplot
- ★ Matplotlib
- ★ Bokeh
- ★ Plotly
- ★ Numpy

II. Data Acquisition & Procedures

Procedures:

1. Scrape data from the following websites:
    a. City of Hope - age ranges and percentages of cancer diagnoses that correspond with each range.
    b. Wikipedia Country Population by Age - extract 2023 table from The World Factbook to use with total population in each country and calculate the totals for each age range (0 to 14, 15 to 64, 65+).
    c. Worldometer - extracted table of countries in the world by population from 2024 to then calculate with population by age. Although the age percentages are from 2023 we moved forward with this table due to there only being a 0.01% and the population overall being stable. Used only the population column.
    d. World Health Organization - table downloaded from the prevalence of insufficient physical data among adults (18+) by country. Provides the (self-reported) percent of population attaining less than 150 minutes of moderate-intensity exercise per week, or less than 75 minute of vigorous physical activity per week (or equivalent). Featuring the age-standardised estimate and the crude-estimate for the columns: both sexes, male, and female, we only used the age-standardised estimate.
    e. Wikipedia Countries by Cancer Rate - extracted table of cancer incidence age-standardised rates (including & excluding non-melanoma skin cancer) of 186 countries.
    f. World Population Review - extracted table of education indices from 2017-2022 of 193 countries.
    g. Wikipedia Countries by Healthcare Systems - extracted table of healthcare coverage percentage and its financing system of 178 countries
2. Extract and clean data to create data frames ##with some AI assistance##
3. Create visualizations accordingly

Problems & Solutions:

1. **Problem:** The age ranges outlined in City of Hope don't match up with the age ranges found for the percentage of that population in each country.

   Solution: We condensed the age ranges from City of Hope to align with the ones from Wikipedia. We mapped the 'Under 20' row as '0-14', combined rows '20-34', '35 to 44', '45 to 54', and '55 to 64' to align with '15-64', and finally the '65 to 75', '75 to 84', and 'Older than 84' into '65+.' Although we lost some percentages for the ages of 14 to 20, we took the risk since they are still being accounted for and made up only a small percentage as seen in the graphics to follow.

2. **Problem:** Size of those estimated to be diagnosed with cancer was incredibly small for those between 0-14 since it only makes up 1% of their population and is not very legible in the visualization

   Solution: Separating the ages 0 to 14 from the main graphic showing 15 to 64 and 65+ allowed us to manipulate the y axis to a smaller count which resulted in slightly better results compared to the original but we note that for future use this could have been done better outside of our scope.

3. **Problem:** Rows from the WHO data on insufficient physical activity by country feature "No Data" for 33 countries as well as used different country names (ex. Netherlands vs. Netherlands (Kingdom of the)).

   Solution: Dropped all rows with 'No data' and created a dictionary to map as many of the country names that were mismatched to match those of the Cancer Rate data table from Wikipedia.

4. **Problem:** The data in the tables we scraped may have inconsistencies throughout since we are trying to scrape data of many countries such as missing values (NaN) or formatting errors (i.e., strings with brackets [] or numbers with %.

   Solution: Before processing and proceeding our codes, we cleaned and preprocessed the data by replacing strings in order to be interpreted as the correct class type so we can perform operations like sorting or plotting.

III. Visualization & Interpretation of Individual Research Questions

Age Groups Affected by Cancers

We know the risk of cancer increases with age but we wanted to see what this meant exactly in relation to all the countries. Taking the percentages of cancer diagnoses of each age group and

the population of each age group from each country, we can visualize approximately how many people are affected.

We began with visualizing what age ranges make up the percentage of diagnoses. Using the information found on City of Hope, **Figure 1** depicts the percentages that we then used in calculating according to the population of people in that age range per country.

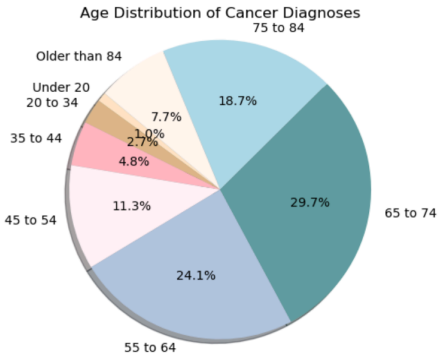

Figure 1

After adjusting the age ranges to follow the format of the age ranges in the country population site available to us, we are left with **Figure 2**. We see that those under the age of 20 (and before adjustments up to the age of 34), do not account for most of the cancer diagnoses.

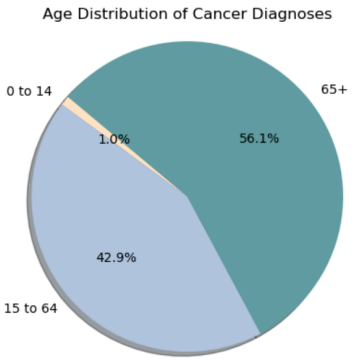

Figure 2

More closely, after estimating the population counts for those between 0-14, 15-64, and 65+ by using the total population for their country and multiplying it with the percentages of each age group, we establish the following table:

	Country	0 to 14	15 to 64	65+	0 to 14 Cancer Cases	15 to 64 Cancer Cases	65+ Cancer Cases
0	Afghanistan	16973701.816	24458336.662	1215453.522	169737.01816	10492626.427998	681869.425842
1	Albania	499725.935	1885278.9045	406760.1605	4997.25935	808784.650031	228192.450041
2	Algeria	13473157.8424	30096918.6132	3244231.5444	134731.578424	12911578.085063	1820013.896408
3	American Samoa	12116.8115	30813.4585	3834.73	121.168115	13218.973697	2151.28353
4	Andorra	10086.5678	55857.1346	15994.2976	100.865678	23962.710743	8972.800954
...	...	...	...	...	...	...	...
197	Venezuela	7138312.9559	18741977.2714	2528093.327	71383.129559	8040308.249431	1418260.356447
198	Vietnam	23671513.5984	69368441.5134	7947730.8882	236715.135984	29759061.409249	4458677.02828
199	Yemen	14260923.8296	24966762.4928	1355477.6776	142609.238296	10710741.109411	760422.977134
200	Zambia	9056724.8044	11674201.4012	584029.7944	90567.248044	5008232.401115	327640.714658
201	Zimbabwe	6347676.7368	9534822.6036	751873.6596	63476.767368	4090438.896944	421801.123036

202 rows × 7 columns

Table 1.1

With our new totals, in order to visualize what makes up the percentages we decided to use bar plots to view and compare by specific country. In our case, we highlight here the United States and select other countries at random to compare: Japan and Sierra Leone. Sierra Leone happens

to be the country with the lowest cancer rate. Below you will see the visualizations to each (broken down in hundreds of thousands), a graph accessible on our Github.

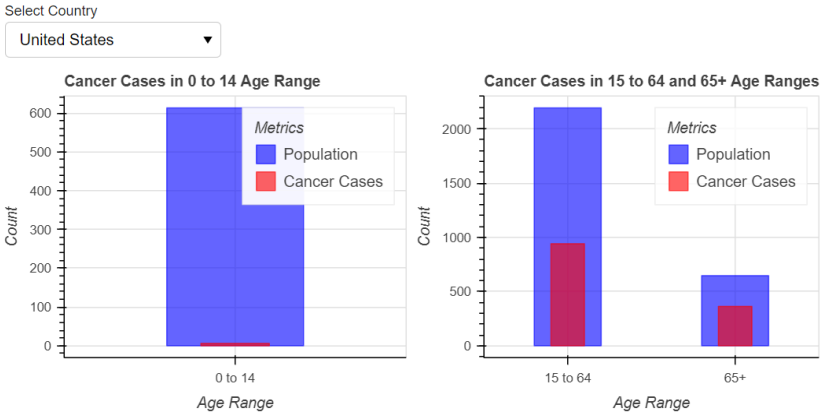

Figure 3

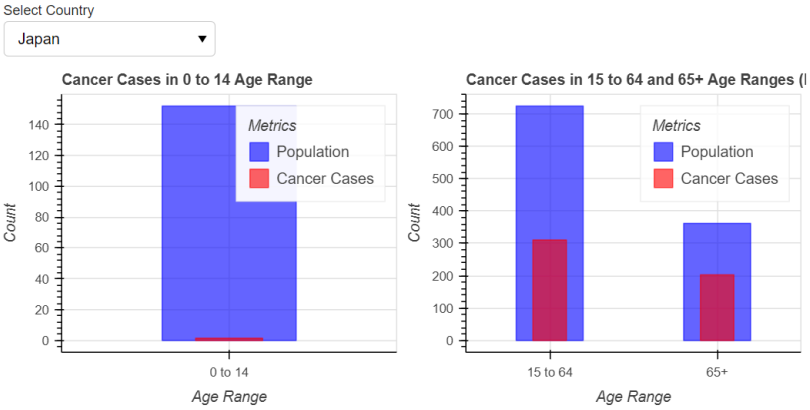

Figure 4

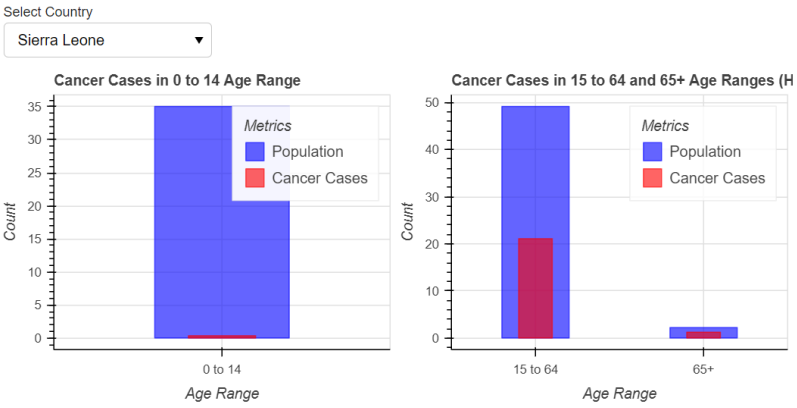

Figure 5

From the visuals (**Figures 3-5**), we are able to see that there are many more cases in the United States across all age ranges compared to Japan and Sierra Leone. Even with an adjusted scale for the 0-14 year olds, their percentage (1%) makes up such a small part of the overall population that despite being hard to read can still be visualized as that those younger than 15 have a smaller chance of being diagnosed with cancer.

Physical Activity Levels by Cancer Rates

With age being an instrumental factor, we also wanted to see how physical activity levels compared to countries and their rates of cancer diagnoses. We hypothesized that countries with lower physical activity levels, in this case represented by the higher inactivity percentage, would be matched with countries with higher rates in overall global cancer incidence (per 100,000).

The United States has numerous weight-based stereotypes that describe the country as lazy and not as active. While this doesn't represent everyone in the country, it is still important to take lifestyles when comparing to cancer rates. From the World Health Organization (WHO), we are able to take each country's percentage of physical inactivity and compare with the country's cancer rate.

First, we cleaned the WHO data and eliminated the columns that were of crude estimate values since the data we are dealing with for cancer rates is age-standardised. With theses percentages of physical inactivity, we sorted them from least to greatest since the smaller the percentage, the more active the country. When scraping the cancer incidence rates we decided to include non-melanoma skin cancer despite its prevalence and low mortality rate because it still affects health. For this case we also focused on the rates columns instead of the numbers of cancer incidences allowing for a meaningful comparison as it accounts for the size.

Plotting the percentages against the cancer incidence rate showed a clear pattern as outlined in the graph below:

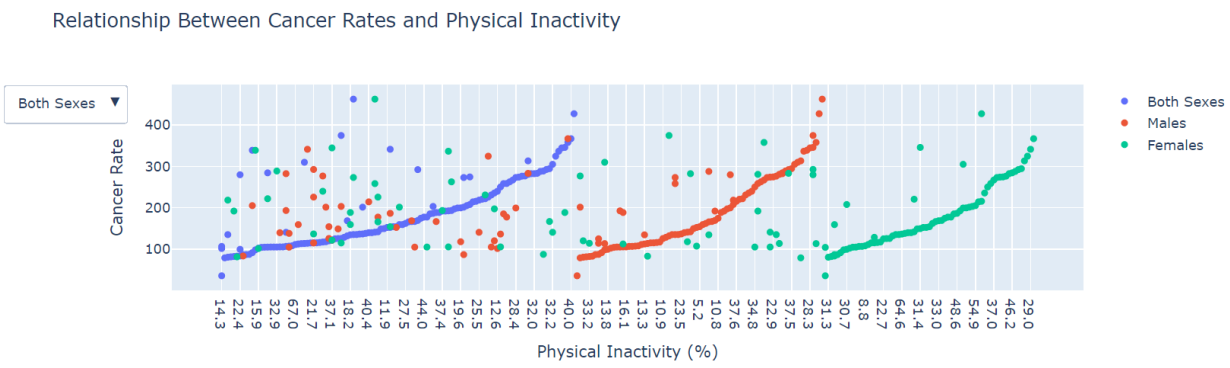

Figure 6: *Screenshot of interactive graph*

Toggle and hover through the different sexes (male and female, or both were only accounted for in this case) and countries on our Github.

Except for a few outliers, it is clear from the **Figure 6** that we were right to conclude that the higher the physical inactivity, the higher the cancer incidence for that country.The outliers would be for further investigation outside the scope of this research problem, For all three figures, the United States was among the highest points on the far right, specifically for women a physical inactivity percentage of 48% was met with a cancer rate of 367. An issue with this case is not tailoring each sex with it's cancer rate which is something for future iterations.

Global Look at Cancer Incidence

To investigate further, we have gathered data on cancer incidence rates and the total number of cancer cases for each country. With cancer being a major health issue worldwide, we want to understand how its incidence varies across different countries in order to identify some potential patterns and factors that may contribute to these differences. The data we gathered displays a comprehensive overview of cancer incidence in 186 countries. This cancer incidence includes both the numbers of diagnosed cases and the rates that they occur per 100,000 people. Our dataset will allow us to examine patterns across these countries and analyze how lifestyle factors, education levels, and healthcare systems may influence and impact cancer incidence. By analyzing the relationship between these factors with cancer rates, we hope to gain valuable insights that could help educate the public and push for potential strategies to intervene.

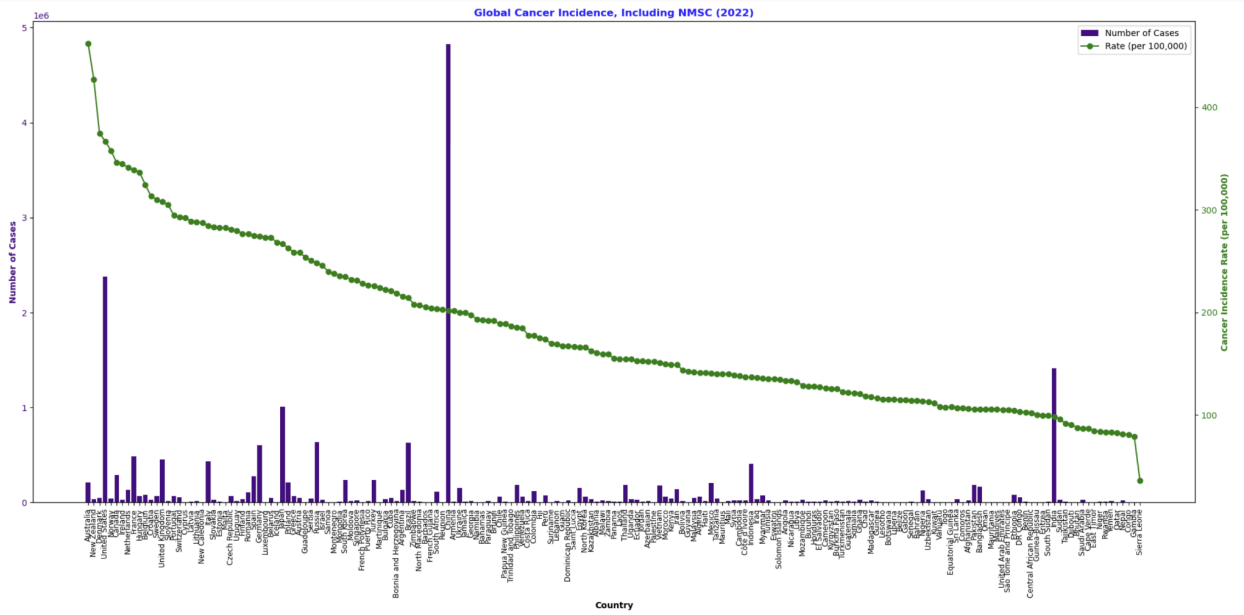

Figure 7

The figure above (**Figure 7**) shows the data we have gathered about global cancer incidence, which includes non-melanoma skin cancer. This displays the total count of diagnosed cancer

cases in each country along with the incidence rate per 100,000 people in 2022. To maintain consistency across all the countries, we will be utilizing the cancer incidence rate as our measurement of cancer occurrence because it accounts for the difference of population sizes and it is age-standardized. We cleaned the data and excluded 'World' from the countries since it posed as an extraordinary outlier and may interfere with our study.

Looking at our plot, we can see that the top four countries with the highest cancer incidence rate are Australia (462.5 per 100,000 people), New Zealand (427.3 per 100,000 people), Denmark (374.7 per 100,000 people), and the United States (367.0 per 100,000 people), despite having a lower number of cancer cases in comparison to China. These countries reportedly have higher cases, which could be influenced by a variety of factors*: education level, lifestyle features, and access to healthcare*. As we continue to investigate, we will explore these countries with high incidence rates to see what factors may be contributing to their elevated rates. We will do this by observing global data of age groups affected by cancer, education level, lifestyle factors, and healthcare systems so we can reveal potential insights to *make a change.*

Education Level

Education plays an important role in public health. We can often observe it increasing awareness of health risks, healthier life choices, and even improved access to healthcare. To look further, we wanted to examine the global Education Index. The global Education Index is part of the Human Development Index (HDI), which is compiled by the United Nations Development Programme (UNDP). It is a measure of education level that is used to rank across all countries, combining factors like average years of schooling and expected years of schooling By looking at the Education Index of the countries worldwide, we can further understand whether higher education levels may correlate with lower cancer incidence rates, more informed cancer prevention and treatment, and better access to healthcare services.

To examine whether there is a correlation, we scraped a data table containing the education indices of 193 countries from 2017-2022. Since this data was gathered over the course of six years, we reflected that in an interactive plot which has been linked in our Github repository.

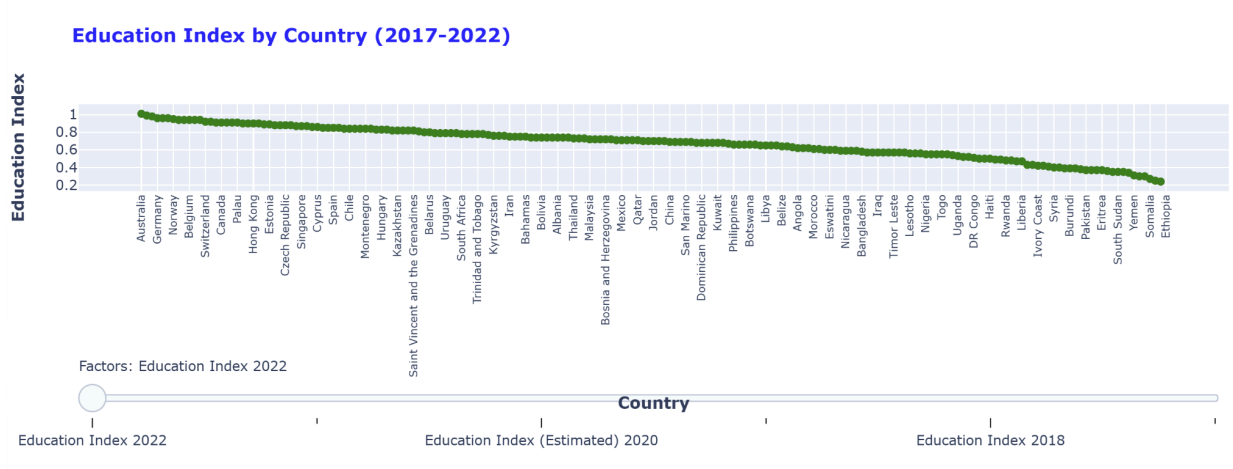

Figure 8: *Screenshot of interactive graph*

Previously, we had noted that Australia, New Zealand, Denmark, and the United States had the highest cancer incidence rates. From this interactive plot, we can observe that these four countries also have some of the highest education indices, which they all are reported as 0.9 or higher consistently throughout the six years. This may indicate that in countries with higher education levels, cancer prognosis is also high because individuals may be more likely to have more knowledge and are better aware of potential cancer risks and preventative measures.

Access to Health Care

While higher education levels in countries like Australia, New Zealand, Denmark, and the United States may influence greater awareness of cancer risks and prevention, healthcare access is also an important factor that could contribute to the high cancer prognosis. To further examine this possible relationship, we examined healthcare systems throughout the world by scraping a data table of health care coverage percentage and its financing system.

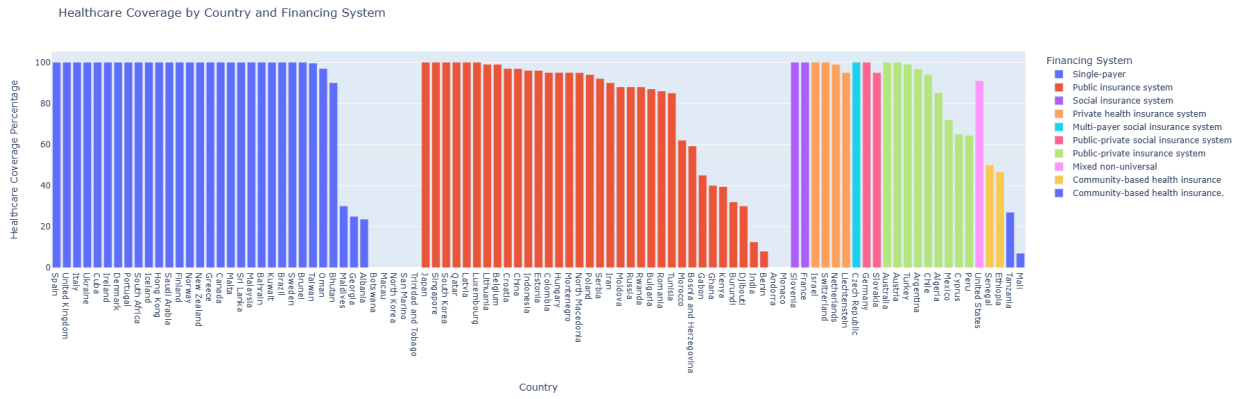

Figure 9

Healthcare coverage and its financing system is critical for determining the effectiveness of cancer treatment and prevention strategies. From this plot (**Figure 9**), we can observe that countries like Australia, New Zealand, Denmark, and the United States have extensive healthcare. They also have a majority of single-payer healthcare where the government funds healthcare to all their citizens despite their income or employment status. This alludes that in countries with more extensive healthcare, individuals are more likely to have access to regular health care services, screenings, and early diagnostic tools which may explain high cancer incidence rates observed in these countries along with higher levels of education. The accessibility of their healthcare systems and financing systems likely contributes to the higher reported cancer rates, as early detection and treatments is more readily available which leads to greater cancer prognosis.

IV. Conclusion

Conclusion

Overall, there are many factors that can contribute to a cancer diagnosis. Here we outlined how age, physical inactivity, education level, and access to health care can relate to cancer.

We recognize that cancer is a sensitive and nuanced subject that ultimately comes down to the individual. Being able to illustrate how important it is to stay active, educated, and have access to healthcare can help in reducing the chances. Although I (Anne) may never get the answer as to why exactly I developed the disease, I can see now that I was part of the small percentage of 20 year olds that were diagnosed and that staying active and educated is key.

In an overview, the correlation between age, physical activity, education level, and access to health care plays a significant role in understanding trends that correlate with cancer prognosis. Our analysis on the global cancer incidence reveals how countries with high cancer rates also tend to have high education levels in which individuals are more informed of health risks and are more likely to take preventative measures such as increasing physical activity. In relation, higher cancer incidence rates have shown to be correlated with healthcare systems in each country where there is more extensive healthcare coverage, making access to diagnostic tools more available. These findings all suggest that the combination of age, physical inactivity, education level, and healthcare access play critical roles in shaping cancer incidence rates worldwide. BY addressing these gaps in the intersection of these factors, countries can work towards reducing cancer rates and help improve public health outcomes country by country.