# Cancer Incidence in the United States by State and Race

Lesrene Browne
Data Science and Psychology

Khushi Morparia
Data Science and Behavioral Neuroscience

Karenna Ng
Biology and Math

**ABSTRACT**

When cancer is detected early during preventative screening, the risk of dying from that cancer is significantly reduced. Our tool will help users understand their personal risk to certain cancers, and could be the first step to push people to ask their doctors about preventative measures. This visualization also incorporates socioeconomic and insurance data, which could be useful for patients when considering the cost of these health procedures. Researchers can also use this tool to identify trends between cancer occurrence and healthcare affordability in different communities. Encouraging early cancer screening and affordable healthcare will help reduce cost of care and mortality from cancer.

## 1 INTRODUCTION

Cancer is one of the leading causes of death in the USA. When diagnosed at an early stage, mortality rates are significantly lower in many cases. One way to help increase the probability of detecting cancer early is to encourage patients in at-risk populations to get screened. As a tool to help patients understand their personal risk to certain cancers, we present a visualization that aggregates and presents information on cancer incidence. Patients can filter the view to be specific to their community, and use that information to help inform whether cancer screening is right for them. This visualization tool could be the first step to bring awareness to people and push them to get screened early and act on preventative measures. The tool will also include information about poverty rates and insurance, which can be useful to researchers looking to determine healthcare needs and affordability in different populations. By encouraging screening and affordable healthcare in the populations that are most at-risk, mortality rates from cancer can be reduced.

## 2 RELATED WORK

While there are many articles analyzing disease data with the focus of identifying trends and biomarkers, there are fewer that are presenting statistics to be accessible to the everyday person. One previous work that did was by Maciejewski et. al [1], who expressed concern for skewed cancer statistics when data collections are small, and proposed an interactive visual analytics tool to better explore and more accurately represent the data. Their tool utilized AMOEBA (A Multidirectional Optimum Ecotope-Based Algorithm) to cluster counties based on a chosen population demographic field. They also illustrated the benefits of displaying data by location and similar statistics over time in the same view, which provides the user with context and enables smarter hypothesis generation. Though our data does not have a temporal element, the idea of coupling different types of visualizations together in the same view seems very effective. In addition, our data is already aggregated by state. While we do not have more granular data, keeping in mind the effects of small sample size on population statistics provides us with more context on the shortcomings of our visualization tool.

Another previous work that is very similar to our goal is by Shenson and Joshi [2], who created a tool called *DiseaseTrends*, that visualizes socioeconomic attributes, such as education level and environmental factors, against diabetes and cancer prevalence in the United States. It is similarly structured as the aforementioned — an interactive choropleth map with small visualizations on the side — but these coordinated views display socioeconomic variables of interest, which is more similar to our goal. They also added interaction to be able to select multiple regions and directly compare their data, which is a useful feature to potentially include in our tool.

## 3 USE CASE

To demonstrate the value and potential of our visualization tool for informing and encouraging at-risk communities to get screened for cancer, we present a usage scenario in which our tool is the first step towards cancer pre-screening. A patient who is interested in learning about their personal risk of certain cancers based on location of residence and race may use this tool to check those statistics. They would be able to see the number of occurrences of a certain type of cancer in their area. They can also use the tool to view insurance and poverty level information, which may help determine the affordability of screening. Another usage case is for a researcher trying to study the effects of poverty and race on the incidence rates of different types of cancer in certain communities. This person could use this visualization tool to see the number of occurrences of a certain type of cancer in their state. Coupled with information about insurance, poverty level, and race, they could use this tool to inform decision making about healthcare pricing and needs in different communities.

## 4 DATA

https://www.kaggle.com/datasets/salomekariuki/cancer-incidence-in-the-us-by-state-and-race.

We obtained our data from Kaggle at the above link. We were able to access three separate CSV files from this source. The first CSV ("Cancer.csv") contains data on the types of cancer and their categories and was originally sourced from the Centers for Disease Control and Prevention (CDC). The second CSV ("Cancer_Occurrence.csv") has information on the occurrence of different cancers (represented by IDs) across different racial groups in all 50 states. This data also came from the CDC. The last dataset ("State.csv") has information on the population of each state and the populations under poverty and insured. It was sourced from State Cancer Profiles. As this data was compiled before we retrieved it, it is missing the years that it represents and links to the original sources; only the name of the organization that collected the data is included. Within the data however, it is very clean and there are no outliers or missing values.

One potential bias of the data can be inferred from the large number of "0 occurrences" of types of cancer in POC race categories. Based on the prevalence of cancer, these statistics are likely not 100% accurate, and may reflect a bias of how the data was sourced by the CDC. Another potential bias is that there are only 5 race categories, and the "Other races and Unknown Combined" category is not inconsequential in size. The source could have been more specific with their demographic data collection and prevented such grouping.

We merged the 3 CSV files by Cancer_id and State_name using Python's Pandas library in Jupyter Notebooks. We also removed cancers that had less than 40000 overall incidences across all states

and racial groups, as the original combined dataset was too large. No new attributes were made, however some are calculated dynamically in our visualization tool, namely incidence ratios.

## 5 Design Process

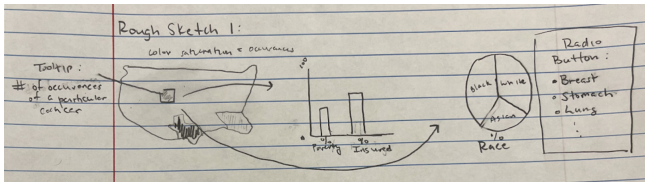


Figure 1: Rough Sketch 1

Rough sketch 1 (Figure 1) shows a map of the USA with a coordinated bar chart and pie chart. On the right is a radio button, where the user can choose what type of cancer to filter the main visualization by. The map is color-coded by cancer occurrences, with a tooltip that displays the actual number of occurrences. When the user clicks on a state, the bar chart shows that state's data about poverty rates and percent population insured, and a pie chart shows cancer occurrence broken down by race. Overall this view is very easy to read, and communicates all of the information in the dataset. Representing information using a map is easy for everyone to read, but state size is not proportional to population, and so some of the percentage statistics may be misinterpreted.

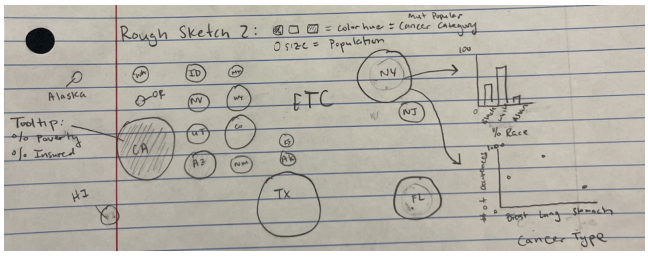


Figure 2: Rough Sketch 2

Rough sketch 2 (Figure 2) shows a map of the USA where each state is represented by a circle proportional to its total population. The color of each state represents its most common cancer type. A tooltip is included to show the percent insured and below poverty in each state when the user hovers their mouse. Coordinated views appear when the user clicks on a state, namely a bar chart breaking down occurrence by race, and a scatterplot showing number of occurrences for different types of cancers in that state. This map view is a little bit harder for the average person to read, but it may be helpful for more accurately understanding some statistics. Also, the scatterplot encoding for different types of cancers was a bit confusing, so we opted to change it for our final sketch.

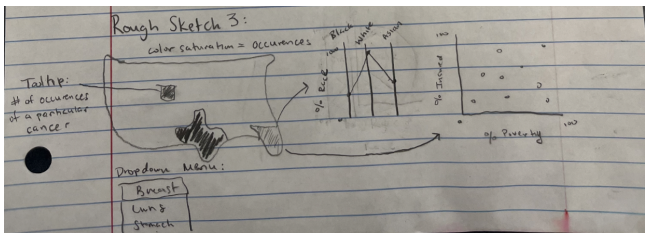


Figure 3: Rough Sketch 3

Rough sketch 3 (Figure 3) has the same map as rough sketch 1, but the coordinated views are changed. Rough sketch 3 also uses a dropdown menu to filter between different cancer types rather than a radio button like in sketch 1. We felt that the dropdown menu would be more aesthetically pleasing than a radio button because it shows less options at one time. The poverty and percent insured data is presented in a scatterplot, which may be more useful in finding patterns between these socioeconomic variables, which aligns closer to the needs of the researcher use case. Occurrence by race is presented in a parallel coordinates plot. While it is easy to compare rates between groups, it is difficult for the everyday person to read so we did not include it in the final sketch.

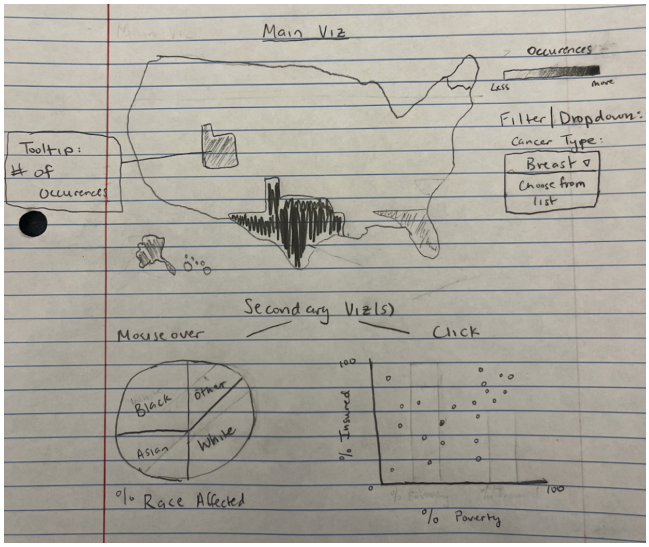


Figure 4: Final Sketch

The final sketch (Figure 4) uses the map and pie chart from rough sketch 1, and the scatterplot from rough sketch 3. The map shows states color-coded by cancer occurrence with tooltip displaying the number of occurrences as calculated using the state's total population. The user can filter the view using the dropdown menu on the right. A user can select a state by clicking on it, and smaller visualizations below the map will show more state-specific information. The pie chart on the left breaks down cancer occurrence by race, and the scatterplot on the right, which shows country-wide data for percent in poverty vs percent insured, will highlight the point corresponding to the selected state.

This final sketch uses visual encodings that are easy to understand, and the interactions are logical but not too busy. We also intend on adding a key for reading the map, and written instruction on how to use the visualization tool. As one of our intended users is everyday people, adding these will make our tool more accessible.

The marks of our tool are areas and points. In the map, areas represent each state. In the pie chart, areas represent the parts of the whole. Points to represent the percent of the population impoverished vs the percent of the population insured in the scatterplot. The channels in our tool are color (hue and saturation), position (vertical and horizontal), shape, and tilt. Color saturation is used to express the magnitude of occurrences of a particular cancer. Color hue represents each race category in the pie chart. Position and shape help to identify each state. Tilt represents the angle that determines how big a slice of the pie chart is which shows the magnitude of each race category with that particular cancer. In the scatterplot, vertical position represents magnitude of the percentage of the population insured, and horizontal position represents magnitude of percentage of the population below the poverty line.

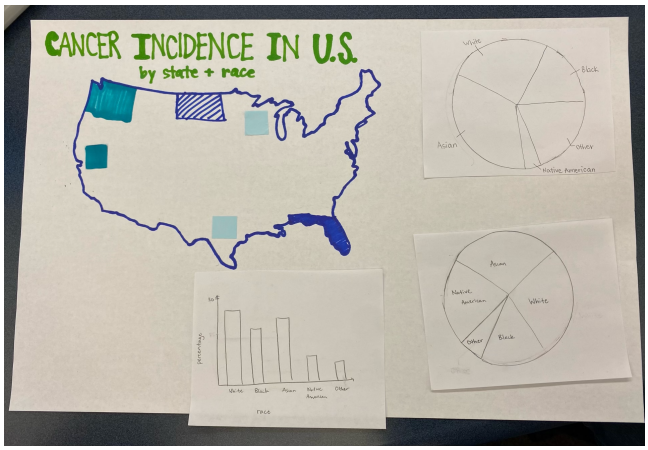


Figure 5: Low Fidelity version of our tool for testing

Before we continued, we created a low fidelity version of our visualization (Figure 5) which only showed our map and different options for a linked graph showing race data for usability testing. We included two versions of the pie chart. One showed race name as a tooltip and the other had the race name in each slice of the pie chart. We also included a bar graph showing percentage where each bar was a racial group.

The feedback we received was that the map was easy to read and understand and the pie chart with the tooltip was the best option shown. Testers also suggested we try changing the pie chart by using a legend.

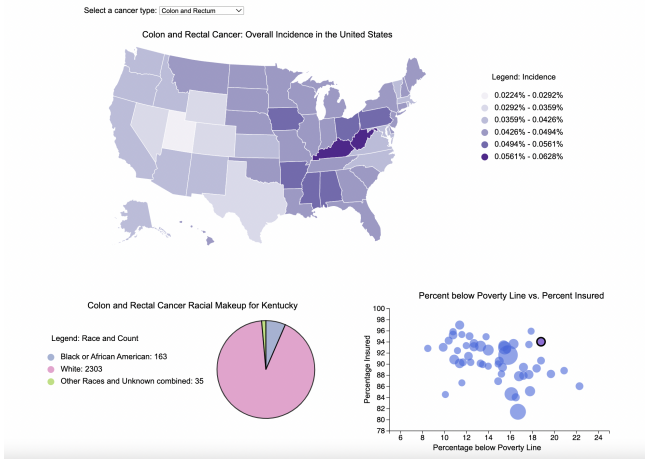


Figure 6: Initial Draft of our tool based on the final sketch

After completing the sketches, we built the tool in parts. First we built the map with its corresponding cancer type filter dropdown, then the scatterplot which was linked to the map by highlighting the corresponding state's point, and finally the pie chart which was linked to the the map by changing based on state and cancer type.

At this point our tool looked like our final sketch, with the map at the top and the the pie chart and scatterplot below next to each other (Figure 6).

We did another round of usability testing using a high fidelity version of our tool. Testers got to interact with the map, the dropdown filter, the pie chart and the scatter plot as they were intended to be used.

The feedback we received was that we needed to make our text more reader-friendly by breaking them up into smaller sections. Testers also suggested we add a paragraph to the pie chart explaining why races with 0 cases are sometimes included and how the chart should be interpreted. They also suggested removing the black outline. Testers also suggested a paragraph explaining how the scatterplot should be interpreted and moving the information in the tooltip from over the map to below or beside the plot. Finally, testers suggested we separate our visualizations better. They felt that the scatter plot was less connected to the other visualizations and as such should be below while the map and the pie chart should be side by side.

After taking these criticisms into consideration, we rebuilt our tool.

## 6 Final Design

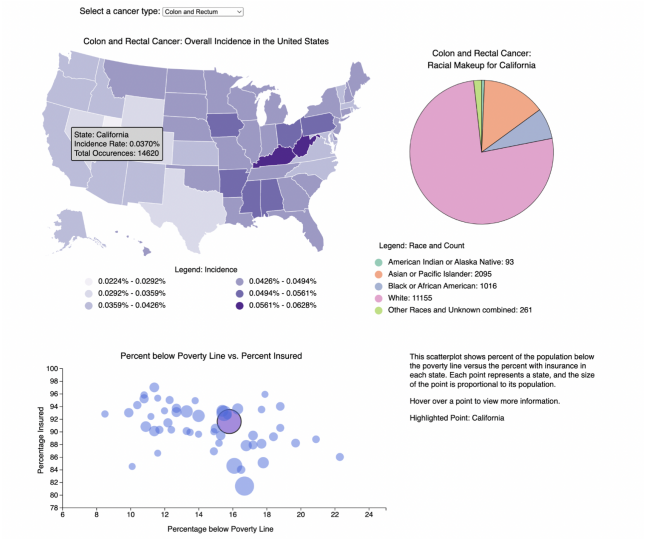


Figure 7: Final Design of our tool

Our final design (Figure 7) is an interactive map showing cancer incidence for a particular cancer. The type of cancer can be changed by choosing a cancer type from the dropdown tool. States are colored by incidence rate, and darker colors mean a higher incidence. The legend below the map shows what percentages the colors correspond to. Hovering over a state shows the specific rate and total occurrences of that particular cancer in that state. Clicking on a state shows the racial breakdown of its cancer incidences in a pie chart to the right. The legend below the pie chart shows the number of cases per racial category. Racial groups that were not surveyed are not shown in the pie chart. Below the map and pie chart is a scatter plot, showing percent insured vs. percent below the poverty line. Each circle represents a state, and the size of the circle corresponds to its total population. The highlighted circle represents the state that was last clicked in the map. Hovering over a circle will show which state it represents, as well as its percent insured, percent below poverty, and total population.

A black, female patient from Ohio deciding whether or not she wants to get screened for breast cancer could open this tool and change the dropdown menu value to 'female breast'. This would bring up the corresponding cancer incidence map. This patient could then hover over Ohio to get information about the rate and instances of female breast cancer in her state. The patient could also press on Ohio to see the racial breakdown of the female breast cancer incidences in her state in a pie chart to the right of the map. These visualizations would give the patient a better idea of whether or not she is at a high risk of having breast cancer and in turn whether she

should seek screening. The patient could also see where her state lies in terms of insurance and poverty levels and consider whether screening would be affordable in Ohio.

A researcher studying social determinants of health and health equity could open this tool and examine the maps associated with each cancer type by changing the dropdown menu value to each cancer type. They could hover over each state to get information about rate and enumeration of incidence. The researcher could also press on any state to see how different racial groups are affected by a particular cancer in a particular state. The state that the researcher selects will also be highlighted in the scatter plot at the bottom, so they can compare the total population, percent insured, and percent below poverty of different states. These visualizations would give the researcher a better idea on whether social determinants such as race, poverty, or insurance status have carcinogenic effects on health status.

## 7 Discussion

Our visualization tool addressed our problem to the best of our ability based on the data we had available and our knowledge of visualizations. The data we used is not completely up to date and also has gaps such as missing race data for certain cancers, which makes the tool ineffective for people of certain races in certain states who are looking for those particular cancers. Also, because of webpage restrictions, we were unable to use the entire dataset and had to remove the cancer categories that had less than 4000 overall incidences. This causes a possibility that the user would not find the cancer type they are looking for in our tool.

A further limitation of our tool is that the racial breakdown of cancer incidence is shown in isolation. It would benefit to also show the racial breakdown of the state's total population, so that it would be clear if the cancer was disproportionately affecting a certain race, or if our pie charts are simply reflecting the demographics of the state's population. Comparing cancer types is also not a possibility with our visualizations. Our tool is designed to show data about one cancer type and one state at a time, resetting each time the dropdown menu is changed. If a researcher wanted to compare data about different cancer types in one state, they would need to record values separately.

If we were to make improvements to our tool, we would use data that was more accurate to the current climate. Since the COVID-19 pandemic, all medical statistics have drastically changed, and data that is not up to date is not reflective of that. We would also ensure that every cancer had data for every race, even if it was of zero incidences. This would make the tool user-friendly for everyone in every case. Lastly, if this is possible, we would attempt to use a platform that enables the use of more data in order to encompass the incidence rates of all the types of cancers documented in the CDC. Again, this would ensure that users in every case and situation would be able to benefit from our tool.

If we could augment the dataset, a further improvement would be to include more demographic data as risk factors. Visualizing cancer by state and race alone is interesting, but it is difficult to find a clear causal relationship. Adding other factors such as age and sex would help the user tailor their filters more specifically to themselves, ultimately supporting the intended use case. Similarly, finding poverty and insurance data on the race or individual level (ex. X% of white people with pancreatic cancer in Massachusetts are insured) would improve our tool, as it would provide more specific information to the user and suggest causal relationships between the scatter plot data and cancer incidence.

## 8 Conclusion

This project aimed to show cancer incidence based on race and state in the US in order to help users determine their personal risk for certain cancer types to increase their likelihood of preventative screening. Early screening is the best way to increase survival of cancer, and this tool was designed to support that process. We cleaned data sourced from the CDC, created an interactive map for each cancer type with linked pie charts showing race data for every state, and scatter plots to show socioeconomic data. The side-by-side scatter plots stand to help users make health decisions that have large cost implications. Although this tool is not a comprehensive look at how all cancer incidence stands in the country today, it can still be used as an entry-level tool for researchers to gain an idea of patterns between cancer occurrence and healthcare costs in different racial and regional groups. It can also be used by cancer-prevention organizations and primary care physicians to help certain populations visualize their risks and motivate them to get screened.

## 9 Acknowledgements

We would like to thank Professor Ab Mosca, for their guidance on the project topic, help troubleshooting code, and for teaching this course. We would also like to thank the members of the "Project Jester" team, for being our test users and providing invaluable feedback on our design.

## References

[1] R. Maciejewski, T. Drake, S. Rudolph, A. Malik, and D. S. Ebert. Data aggregation and analysis for cancer statistics - a visual analytics approach. In *2010 43rd Hawaii International Conference on System Sciences*, pp. 1–5, 2010. doi: 10.1109/HICSS.2010.128

[2] J. Shenson and A. Joshi. Visualizing disease incidence in the context of socioeconomic factors. In *Proceedings of the 5th International Symposium on Visual Information Communication and Interaction*, VINCI '12, p. 29–38. Association for Computing Machinery, New York, NY, USA, 2012. doi: 10.1145/2397696.2397701

## 10 APPENDIX

### 10.1 Data Abstraction

Items (rows)

- Each item represents the number of individuals with a particular type of cancer in a racial group in a state

Attributes (columns)

- Cancer_id: ordered, quantitative, sequential
- Cancer_type: categorical
- Cancer_category: categorical
- State_name: categorical
- Race_name: categorical
- Count: ordered, quantitative, sequential
- Percentage population below poverty in state: ordered, quantitative, sequential
- Percentage population insured in state: ordered, quantitative, sequential
- State_population: ordered, quantitative, sequential

### 10.2 Task Abstraction

Domain task: patient learning about their personal risk of cancer

- Actions
  - High level: Analyze - Consume - Present
  - Medium level: Search - Locate
  - Low level: Query - Identify
- Targets
  - Attributes - Many, Similarity

Domain task: researcher comparing cancer statistics between specific states

- Actions
  - High level: Analyze - Consume - Discover
  - Medium level: Search - Explore
  - Low level: Query - Compare
- Targets
  - Attributes - Many, Similarity