# A Study of Cancer Mortality Incidence Ratio in the US from 1999 to 2018

Sam Lindsay, Peter Xu

March 15, 2022

## 1 Summary

1. Which U.S. States has had the greatest percent decrease in their cancer Mortality-Incidence Ratio (MIR)?

   The state with the greatest decrease was Tennessee, with a decrease of 32%.

2. Which type of cancer have seen the greatest decrease in their Mortality-Incidence ratio?

   The greatest decrease was cancer of the floor of the mouth, which had a 100% decrease.

3. Within the same county, how does the Mortality-Incidence Ratio differ between races?

   The data was incomplete, but almost uniformly white people had lower Mortality-Incidence Ratios.

## 2 Motivation

In recent years, there is a growing recognition of social factors affecting individual and public health performance. In other words, scientific communities are realizing the impact of social status on people's lifestyle choices, which subsequently drives people away from a healthier life. Through this project, we hope to provide some future study topics by pointing out the differences in cancer treatment outcomes among people in different states, or different ethnicity. This data may be helpful for anthropologists and policy makers to better direct research funds such that the priority is better organized.

## 3 Dataset

The CDC publishes tables that have aggregated data relating to the incidence and mortality rates for types of cancer. Each table is broken down by race, by gender, and by the type of cancer, amongst other things. Furthermore, the data set is split into multiple different tables, with tables breaking all of the data down either by age, state, or county, amongst other things. The tables can be accessed at the following link:
`https://www.cdc.gov/cancer/uscs/dataviz/download_data.htm`
Particularly, we are using the files from 1999-2018. We also made use of geospatial data from the U.S. Census Bureau in order to get accurate maps of each county in the United States.
`https://www.census.gov/programs-surveys/geography/technical-documentation/naming-convention/cartographic-boundary-file.html`

## 4 Methodology

1. We clean the data by removing rows with `NaN` values, specifically ' ', '+', "-', and '.' values in numeric columns.

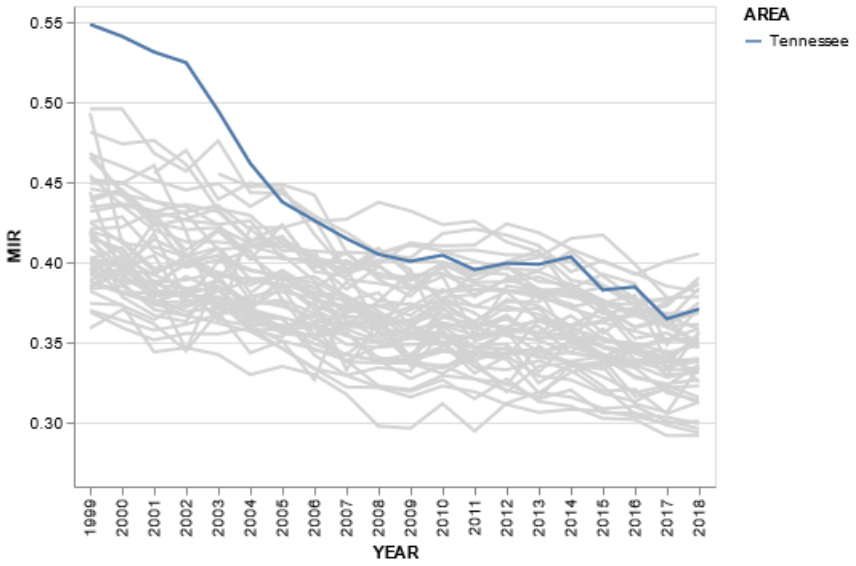
2. Then we calculate the MIR for all datasets, using the incidence rate to divide incidence rate.

3. For the first research question:
   (a) We only include the rows that are representing data for all races, all cancer types, and all sexes combined in our analysis. This is our design choice to achieve maximum coverage of the real population.
   (b) We group the data by each state.
   (c) Then, we compare the earliest and latest year with a valid mortality-incidence ratio in the data set to find a percent change.
   (d) Now, we compare each state in order to find which has the greatest rate of change.
   (e) Lastly, we use Altair library to create an interactive line chart. Viewers can select a specific state and see the change of MIR in this state from 1999 to 2018.
4. For the second research question:
   (a) Similar to the previous question, we only look at the rows that represent data for all races and all sexes. This is also for the purpose of covering the most portion of population.
   (b) We group the data by each type cancer.
   (c) Then we compare the earliest and latest years to get a rate of change for the MIR.
   (d) Now, we compare each cancer type to see which has had the greatest drop in MIR.
   (e) Lastly, we use Altair library to create an interactive line chart. Viewers can select a specific cancer site to see the change of MIR in this state from 1999 to 2018.
5. For the third research question:
   (a) We use the table that is broken by county, we first filtered down to rows representing all cancer sites and all sexes. Also, Alaska and Hawaii are excluded from the visualization for clarity purpose.
   (b) We extract "geoid" from the area information in filtered dataset, which is the unique identifier in the shape file.
   (c) Then, using geoid from both the DataFrame and shape file, we join them together using left join on shape file.
   (d) Lastly, we use Altair library to create an interactive map. Viewers can select a specific race to see the difference of MIR among all states except Alaska and Hawaii. This enables an easy comparison between different races.

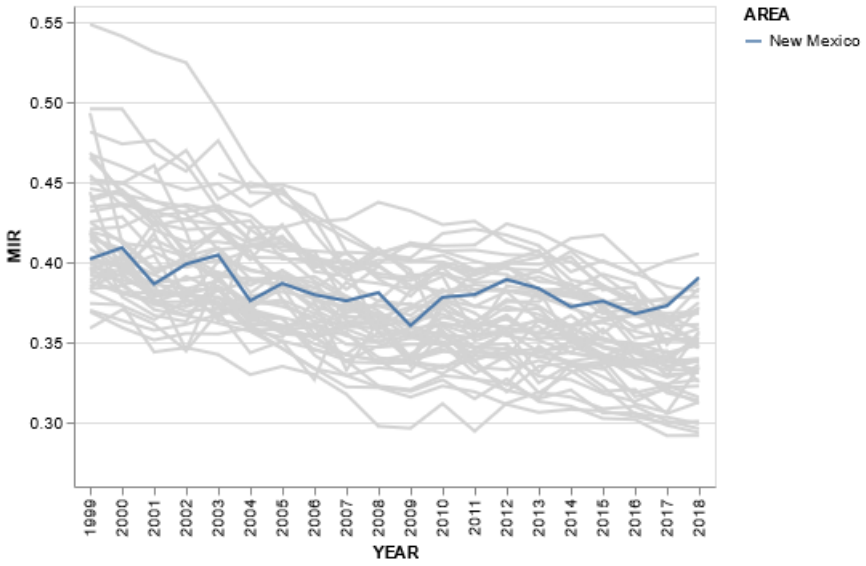
# 5 Results

## 5.1 Which State has had the greatest decrease?

Our result is that Tennessee had the greatest decrease in MIR over the time period of 1999-2018. This decrease corresponded to a percent change of 32%.




Tennessee's decrease is impressive, but it is perhaps more interesting to note that there was a general trend downward for all states that are tracked in this data set. As a result, it may be more interesting to note that New Mexico had almost no change over this time period, having a decrease of only 2% in the MIR for all cancer types combined.




This chart can be interacted with to explore all of the states at the following link:

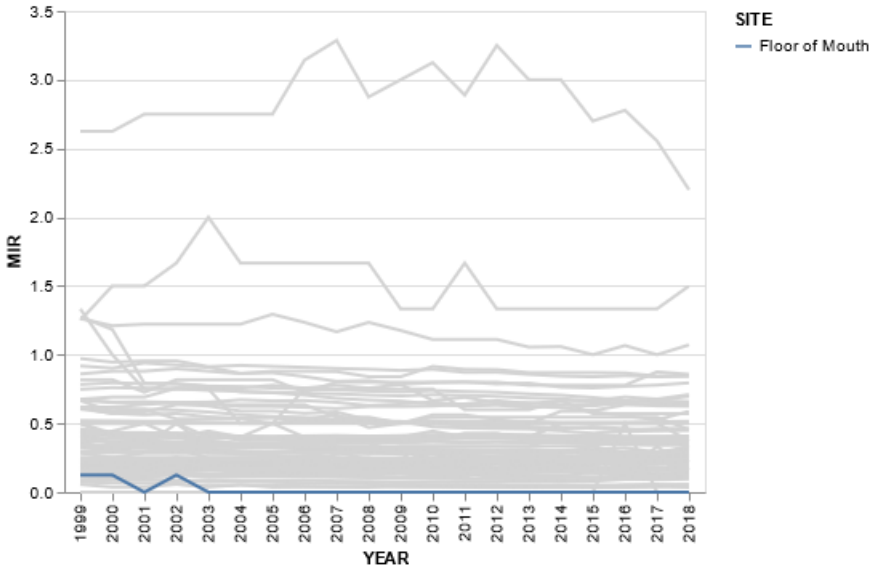
`https://samuel-lindsay.github.io/cse163-cancer-research/state_improvement_plot.html`

These results point towards a general trend downwards in the Mortality-Incidence rate of all cancer types combined. This may be the result of a few different things. The first is that testing has improved,

which would increase the incidence rate of a type of cancer. Increased incidence would decrease the MIR. The second is that treatment has improved, which has led to a decrease in mortality. It may also be some combination of these two possibilities. Moreover, the overall decreasing trend may point at the diminishing of chemical cancer-causing pollutants around the country. And the variation of this decreasing rate can be explained by different public health policies in different states. Some state may invested more effectively in health system than another.

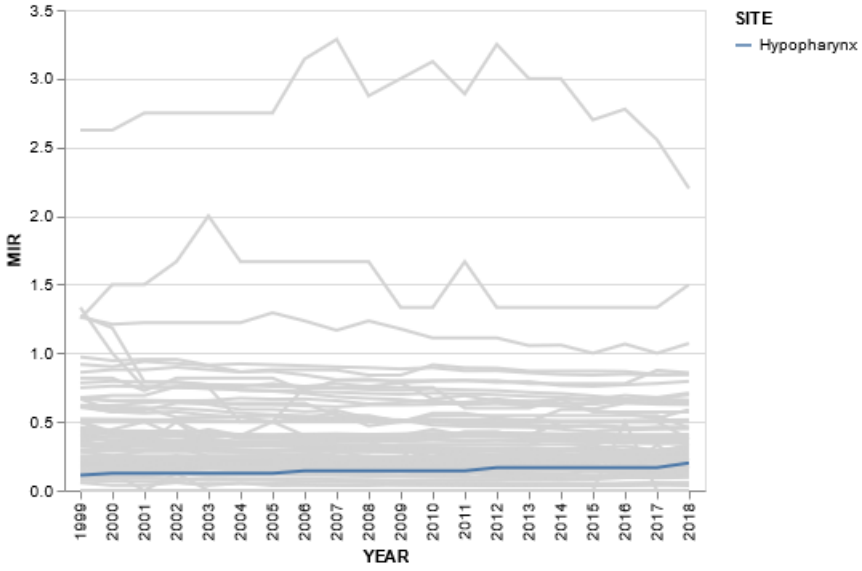
## 5.2 Which type of cancer has had the greatest decrease?

The cancer type with the greatest decrease in its MIR was cancer of the floor of the mouth, which had a 100% decrease.



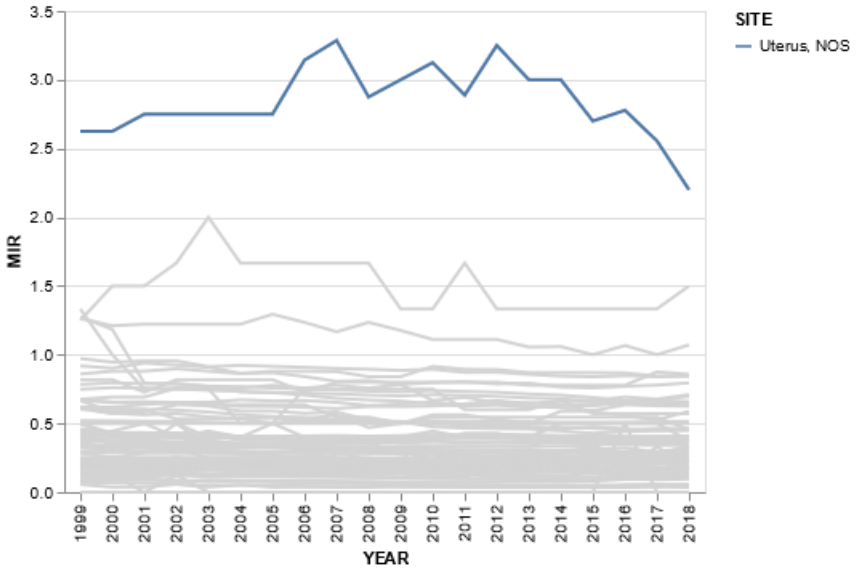

This seems to primarily a result of how rare the cancer is to begin with, as cancer of the floor of the mouth had very few cases over all. Furthermore, all mouth cancers are associated with HPV as well as with tobacco. Tobacco use has largely gone down in recent times, and the first HPV vaccines came out around 2006. These two risk factors going down have likely contributed to a somewhat rare cancer having a mortality rate of 0.

On the other hand, the MIR of hypopharanx cancer has steadily increased over the past 20 years.




This trend is somewhat unique in the data set, and might point to any number of things. As we are not public health students, we do not want to make any strong prescriptions on why this might be the case, just to note that it is the case.

However, the most shocking part of this data may not be the increases or decreases, but the stunning outliers in the data set. Uterine cancer has the highest MIR by far.




Uterus cancer is somewhat known for being difficult to diagnose. As a result, we believe this high MIR is a result of a very large number of people only finding out that they have uterus cancer very late in its development. On a more positive note, this MIR has one down substantially over the time period from 2011-2018, which points towards an improvement of some kind in screening for cancers.

This chart can be interacted with to explore all of the states at the following link:

`https://samuel-lindsay.github.io/cse163-cancer-research/state_race_map.html`

### 5.3 What is the difference in MIR between races?

The interactive map can be seen at the following link:

`https://samuel-lindsay.github.io/cse163-cancer-research/state_race_map.html`

To be honest, the mostly gray map for any race other than white is truly shocking to us. Note that Colorado, Minnesota, and Kansas are missing MIR data completely for all races. Initially, we believe this is due to the small population size of ethnic minorities, but CDC provide detailed lists of counties or states opting out for some report due to the small number. However, Nevada, Minnesota, Kansas, and most of the gray counties are not among the list. Consider the implementation described in Methodology section, we must have both mortality rate and incidence rate for a region in a specific year to obtain the correct MIR. Thus, the large scale of missing data raise a significant alarm about the efficiency of data reporting in the US. Further investigation is needed to answer other questions. With the limited explanation provided by CDC, we can only go this far.

With the data we did have, however, there are a few main conclusions. Race definitely plays some factor in the MIR for all cancer types. The most striking difference is in Native Americans in Louisianna. One region has a significant darker color for them compare to all other races. For African American, the color seem to have a darker fade than other races around the country. However, Hispanic is an exception. We can observe a slightly lighter color in Florida for Hispanic than White population. It is possibly due to the high concentration of white retirees in Florida, while the Hispanic population make up could be younger.

## 6 Impact and Limitations

Many of the impacts of this data have been discussed above in the results category. Broadly, our results might give public health professionals information on where their focus should go in order to improve health outcomes. Furthermore, our results raise questions for future research. For example, why did Tennessee experience such a decrease in its cancer MIR while New Mexico stayed the same? Why has there been a steady increase in hypopharynx cancer? What causes uterus cancer to have such a high MIR, what caused the sharp decrease in its MIR from 2011-2018, and what can be done to continue this trend? Finally, the difference between racial groups points firmly to some sort of racial bias in the healthcare system that should be researched. Almost all Americans know someone who has cancer or has had cancer some time in the past, so the results of these studies have the potential to impact all Americans. Additionally, our research at definitely points to the fact that certain racial groups need more attention from our medical system, and we hope that this research could help to benefit these people.

The impacts listed above implicitly show the limitations of our analysis. Nothing in our data gives solutions nor explanations. The MIR is an inherently limiting metric to use. An increase in MIR is necessarily bad: it must correspond to an increase in mortality or a decrease in incidence without a corresponding decrease in mortality. However, a decrease in MIR is far more difficult to interpret. It could be that the mortality rate of a cancer has dropped, which is good. However, it more often is the result of a change in incidence rates. For example, an improvement in diagnosing existing cancer, or a net increase in the number of people who are getting cancer in a population. This data set tells us nothing about this without further, more in depth research.

Finally, there are limitations and biases present that were discussed in section 5.3. The data on cancer rates of different racial groups is extremely lacking for large amounts of the country, and this makes us unable to speak definitively on anything relating to race. Also, our design choice of excluding Alaska and Hawaii in our visualization for higher clarity can be harmful to Alaskan Natives and Pacific Islanders since they logically have a higher make up in the excluded states.

# 7 Challenge Goals

The selected data set lends itself to two of the challenge goals quite naturally. The first is to use a **new library** in the course of our research. Specifically, we want to create interactive data visualizations to make it more natural to compare different states or different types of cancer. This was done by using Altair

The second is joining **multiple datasets**. Initially, we believe that because the complete data set is broken into multiple tables, it will be necessary to join these separate tables. Throughout the course of our research, we determined that this was not necessary. However, we forgot that in order to make our visualization, we would need to join our data with geospatial data that describes the boundaries of counties and states in the United States. Furthermore, we calculated the MIR by splitting individual data into two separate tables, and then merging them together in order to convert long form data into wide form data that is is possible to calculate the MIR for.

# 8 Work Plan Evaluation

Our work plan was split into five sections.

1. First we were going to clean the data. We budgeted 2 hours for this. We spent over two work sessions at least 8-10 hours on just cleaning the data.

2. Second we made a function to calculate the MIR. We budgeted up to 5 hours for this, but this took just under an hour.

3. The third step was using the calculated MIRs to get the percent changes for questions 1 and 2. We budgets 2 hours and this also took just under an hour.

4. The final step was creating data visualizations. We budgeted 10 hours for this, and it took upwards of 25 hours to complete between both of us working on it.

It is safe to say that our initial estimates were off in each category by a lot. The data was far more difficult to work with that initially expected. The data set is extremely large, so while an initial scroll through the data seemed promising, it hid many problems in the data. On the other hand, after the data was clean, writing the function to get MIRs and then using the MIRs was extremely fast. It took some thinking to create elegant solutions, but it turns out that pandas does a lot more of that work for us than expected. Finally, the visualization estimate was way off for a few reasons. One is simply that we chose to make interactive visualizations for all of our data, instead of just the third question. A second is that it is difficult to know how long using a new library could take - an infinite number of unforeseen pitfalls could come up! Finally, in the process of writing this method, we often found that the data needed to be filtered in a more specific way or grouped in a different way, causing this section to cause (major) rewrites of other sections of code.

# 9 Testing

Our code was split into two main modules, one that cleaned and processed the data and one that created the visualizations. The visualizations were tested by manually interacting with the filters and ensuring that the results corresponded to the correct filtered data. The main testing was done for our data cleaning and processing. Specifically, we tested our methods of filtering data and getting the MIR. For all of these, we used smaller data sets that were specifically chosen to isolate the exact features we wanted to test. The data filtering was tested by using assert statements that compared 'manual' filtering to the intended filtering from our utility methods. The MIR data was tested by using our methods to calculate the MIR and comparing it against MIRs that were calculated by hand (manually dividing a year's mortality rate by its incidence rate).

Since we were able to ensure that the data is filtered properly and that the MIR is calculated correctly for a given data set, we can be sure that the results above are accurate. This is because all of the results above are directly derived from the MIR.

## 10 Collaboration

During Peter's software set up process, he consulted both Hunter and a number of TAs. Also, we consulted Peter's public health professor James Pfeiffer, who helped us understanding the mortality and incidence rates essential to our calculation.