Examination of United States Cancer Mortality from 1999-2017 and trends between Race, Ethnicity, Age Group

Examining Cancer Mortality Through a Variety of Categorical Variables

Brent Cameron

2021-12-24

# 1 Summary/Abstract

With an increasingly aging worldwide population, Cancer and its disease burden has become ever more important in the global sphere. In America in particular, where total cases from 2010 to 2020 were projected to increase by 24.2% [1], increasing knowledge of possible interventions for the disease is of utmost priority. Cancer effects populations nationwide by being responsible for uncontrollable cell growth and metastases, as well as long-term decrease of life quality as a result of side effects of long-term care including chemotherapy. Cancer is also a disease that has established itself with disparities in several types of cancer being more highly concentrated to certain minority groups at a higher level than overall populations [2]. In this research project, data will be analyzed to further understand the trend of cancer mortality trends from 1999-2017 in the United States using data collected via the CDC wonder website. The goal of this project is to understand the role that age group, sex, ethnicity (characterized as hispanic or non-hispanic), and race play in the trend of overall rates of the disease nationwide, as well as to shed light on population groups that are of a higher need for directed health intervention. This research will allow for a greater understanding of populations that can experience the greatest positive impact from targeted health intervention.

# 2 Introduction

## 2.1 General Background Information

The word cancer has long been one of the most recognizable and fear-inducing words worldwide in the sphere of chronic disease. Cancer, along with heart disease and diabetes, are responsible for the leading causes of death and disability in the United States. In 2012, there were an estimated 14.1 million new cancer cases with 8.2 million cancer deaths worldwide [3]. Cancer also affects the elderly at a hightened rate,being the fourth highest cause of death for the very old,with people aged 85 and older making up 6% of total cancer cases [4]. Additionally, cancer disproportionately affects some racial and ethnic groups more than others, with Black patients being more likely to have metastatic disease at diagnosis as well as (along with the Hispanic population) being less likely to receive medical treatment [5].

This disproportional effect to some races over others can be seen in a variety of different types of cancer. For instance, lung cancer has “racial and ethnic differences which in turn impacts the overall epidemiological differences…” [6]. All is not depressing however, as research states that while there are some discrepancies in how cancer effects different groups (with under representation of the Black population in clinical trials being one example), the use of the patient navigation model (where an individual is trained to offer support for patients in trials), has made a positive impact in improving health care access [7].

Because of the vast scope and many different types of the disease, resources and networks have been continually strained in identifying the most efficient and effective ways to counteract cancerous cells. While a multitude of research has already been conducted on cancer, understanding the differences in how the disease continues to affect and interact with different populations can serve as an invaluable resource in tailoring new innovative strategies to combat the disease.

## 2.2 Description of data and data source

The data used in this project is a compilation of information regarding cancer mortality cases from 1999-2017 based on age group, sex, ethnicity. race, and cancer sites (*note that this data only includes information collected from the United States*). The data is composed of 8809 observations and was retrieved from the CDC wonder website which can be found here:

\*<https://wonder.cdc.gov/cancermort-v2017.HTML>

The data was retrieved by using “cancer sites” as the first group in the “group results by” box, while the four aforementioned variables (age group, sex, ethnicity, and race) were placed in each of the four other available “and by” boxes.

## 2.3 Questions/Hypotheses to be addressed

Cancer has been and will continue to be widely studied- however, with the presence of constantly evolving data and new research shedding light on innovative techniques and approaches to better manage and control the disease, new examinations of data trends can provide a valuable framework for future studies. With that being said, there are several questions in particular that I will ask and attempt to answer throughout this research project:

*Are cancer mortality trends dependent on or affected by race in the United States? If so, in what ways?*

*How does age factor into mortality trends of cancer? Will the past research concluding that the elderly are the most vulnerable continue to be held evident?*

*Does ethnicity play a role in either further susceptibility or resilience to cancer (more or less cancer deaths for certain ethnic groups)? If there is a difference between trends for certain ethnicities in the overall disease burden why would that be the case? (E.G. Presence of a good support system or lack thereof, high religious affiliation, healthier overall living, etc.)*

*Does the particular cancer site have an affect on the outcome of the disease? If so, what sites have the largest mortality relative to the number of cases?*

In addition to these supplementary questions, our main hypothesis will focus around the potential role that race and ethnicity plays in cancer rates, and in particular, that there is a measurable and distinct effect on cancer rates as a result of race and ethnicity.

These aforementioned questions will serve as valuable starting points that, if possible, will allow for deeper data analysis to determine synergy between variables (do certain races or ethnicities have a higher propensity for cancer in certain sites? Are there trends between sex and age group relative to cancer mortality trends?). While some of these questions posed may not be able to be fully explored and answered with the data set utilized here, further research in topics of interest will elucidate these questions and allow for a greater reinforcement of already established research and data analysis regarding cancer and its disease burden.

# 3 Methods and Results

The data was gathered from the CDC’s Wonder website, which allows for personalized public health information depending on what’s desired. As the research in this project focuses on sex, age, ethnicity, and sex, those are the four variables that were chosen to compile the original data set, in addition to a few miscellaneous variables that were deemed superfluous and removed, these included mainly ID numbers that corresponded with the data retrieved. In addition, the original data was cleaned by removing all NA variables from the data set after selecting the desired variables for the research project. The selected variables include: Age, Sex, Deaths (Cancer), Ethnicity (Hispanic or Non-Hispanic), Race, and Cancer Site.

## 3.1 Data aquisition and cleaning

As previously stated, the data for this research project was retrieved from the CDC Wonder database. This database was chosen due to the reputable source of the Center for Disease Control, as well as the high quality of reporting and information reported that can be depended on for its reliability and authenticity.

The data had several features removed for sake of more pertinent analysis, as well as the number of observations decreased from 8809 to 8752 due to removal of all NA variables.

## 3.2 Full analysis

Cancer death rates were assessed to determine any statistical significance relating to Race and Ethnicity, utilizing a comparison between a null model and 1. Single Tree, 2. LASSO, and 3. Random Forest models.

All variables except for the Deaths (cancer) variable were categorical. Deaths (cancer) was the main measurable outcome, while Race and Ethnicity served as the two main predictor variables. Five fold cross-validation was used to ensure accurate results. R version 4.1.1 was used for the project.

# 4 Results

Results were inconclusive, although the best performing model (single tree model) initially performed better than the null with an RMSE of 19904, which was below the null model RMSE of 23678, fitting with additional data provided a worse RMSE of 23820, leading to the conclusion that overfitting was present in the model. The LASSO model performed with an RMSE of 21926 and the random forest model performed with an RMSE of 20693, neither of which were adequate for accurate future fitting.

The variables “Non-Hispanic” in Ethnicity, “White” in Race, and “All\_Cancer\_Sites\_Combined” in Cancer Sites were the three variables with the largest measured value of importance in the model. Due to issues with overfitting in the model, it is possible that these variables are also responsible for the problems with overfitting.

The single tree, LASSO, and random forest model plot results have been reported below, along with a summary of the most important variables:

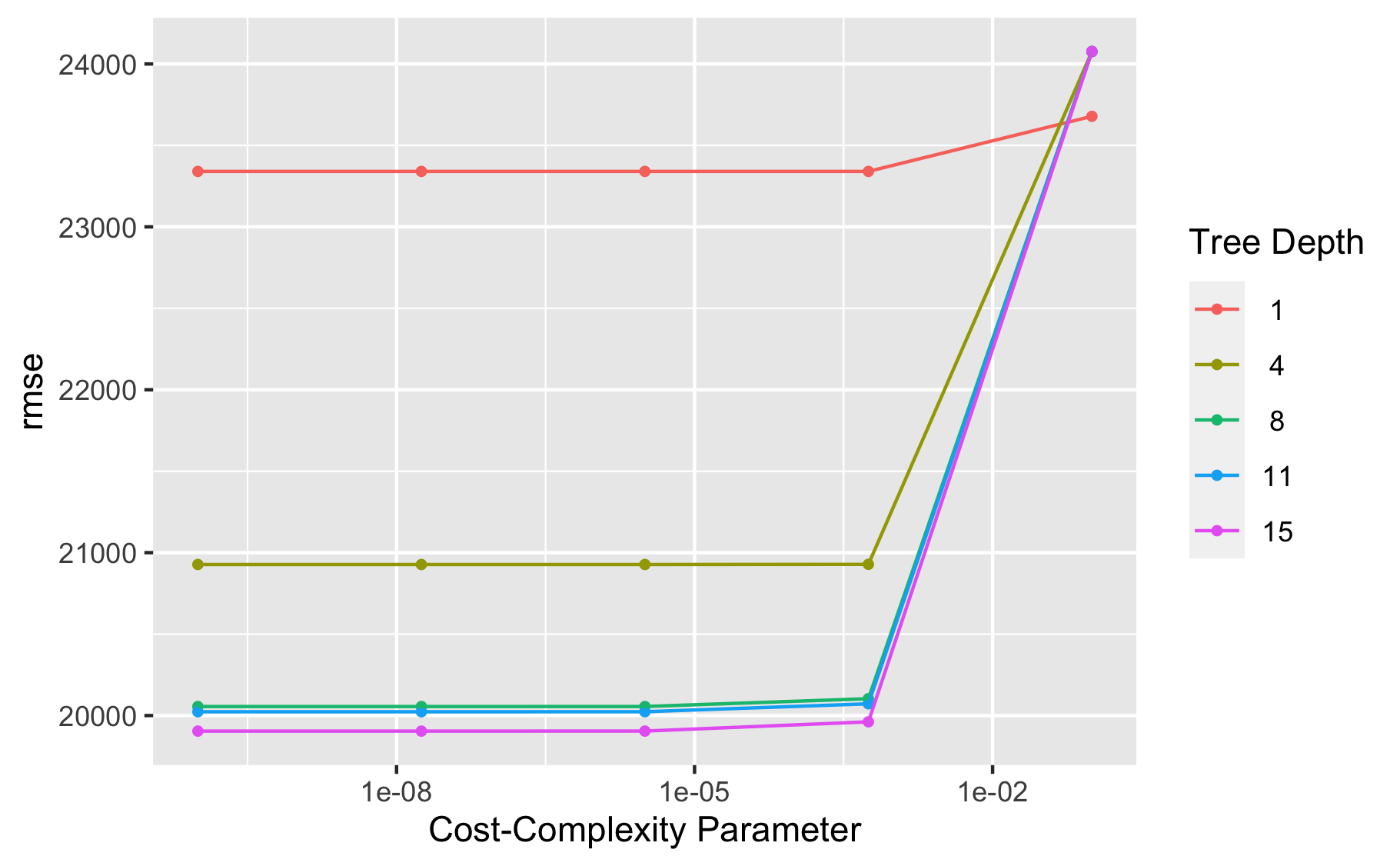


Figure 4.1: Single Tree Plot

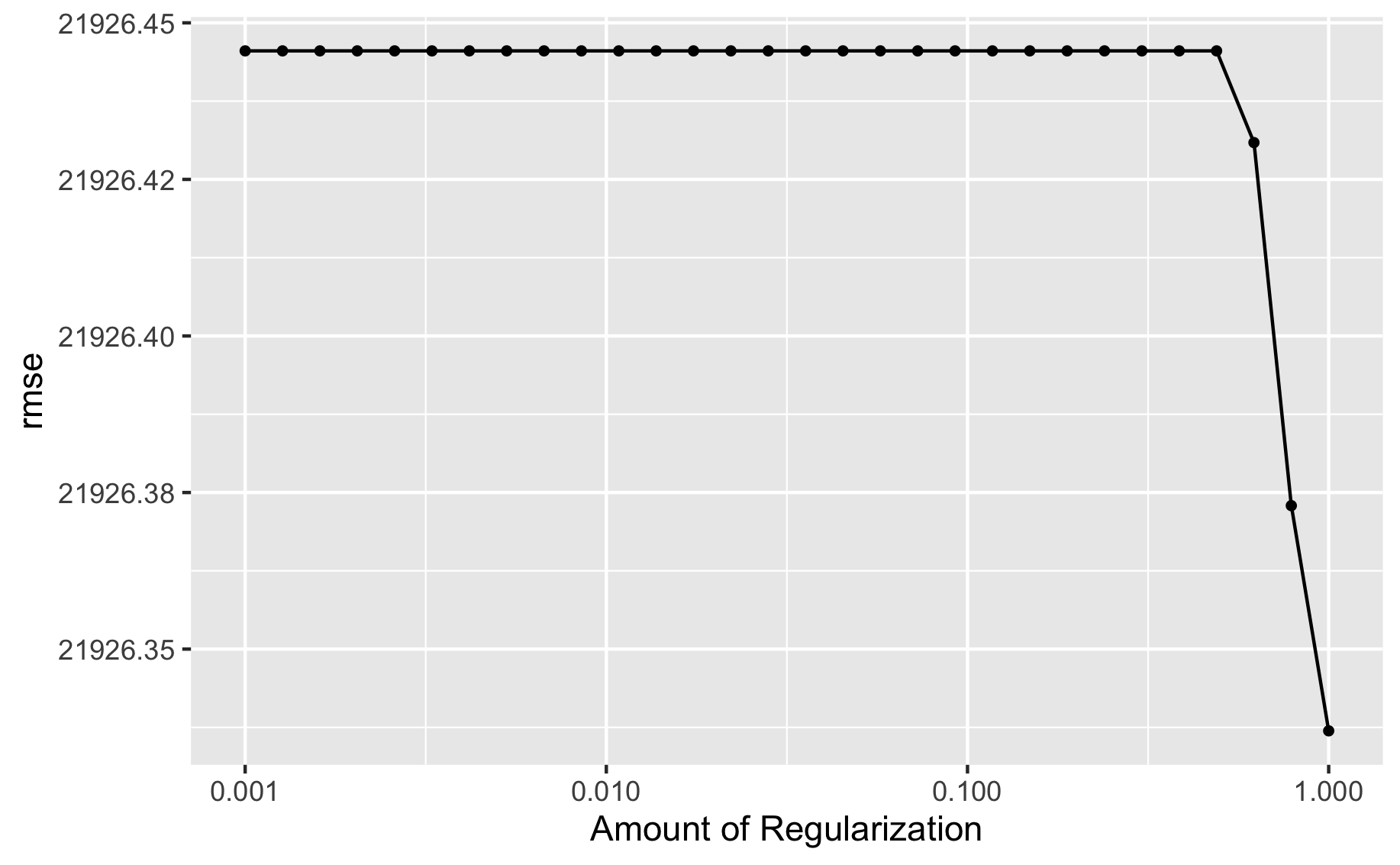


Figure 4.2: LASSO Model Results

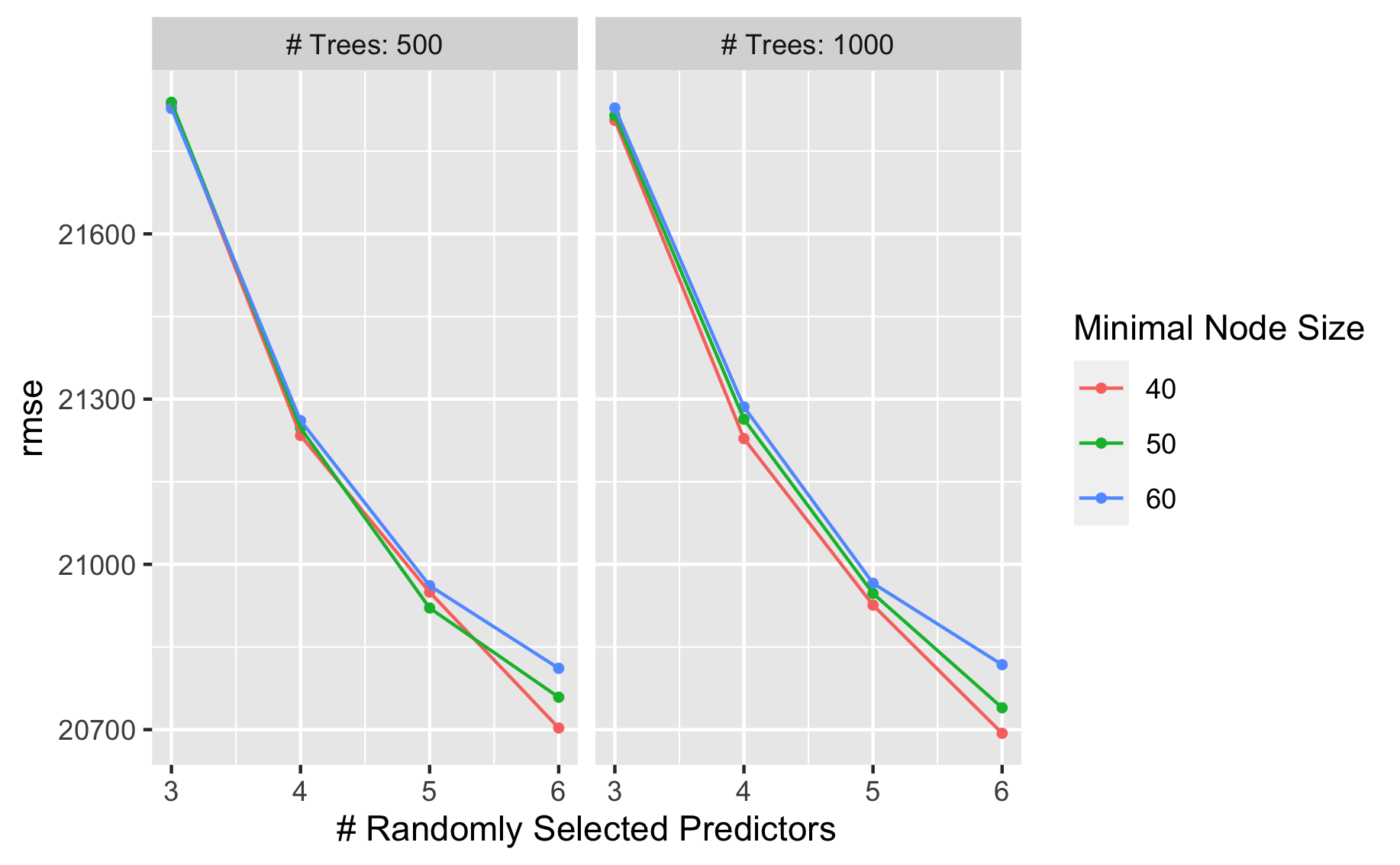


Figure 4.3: Random Forest Model Results

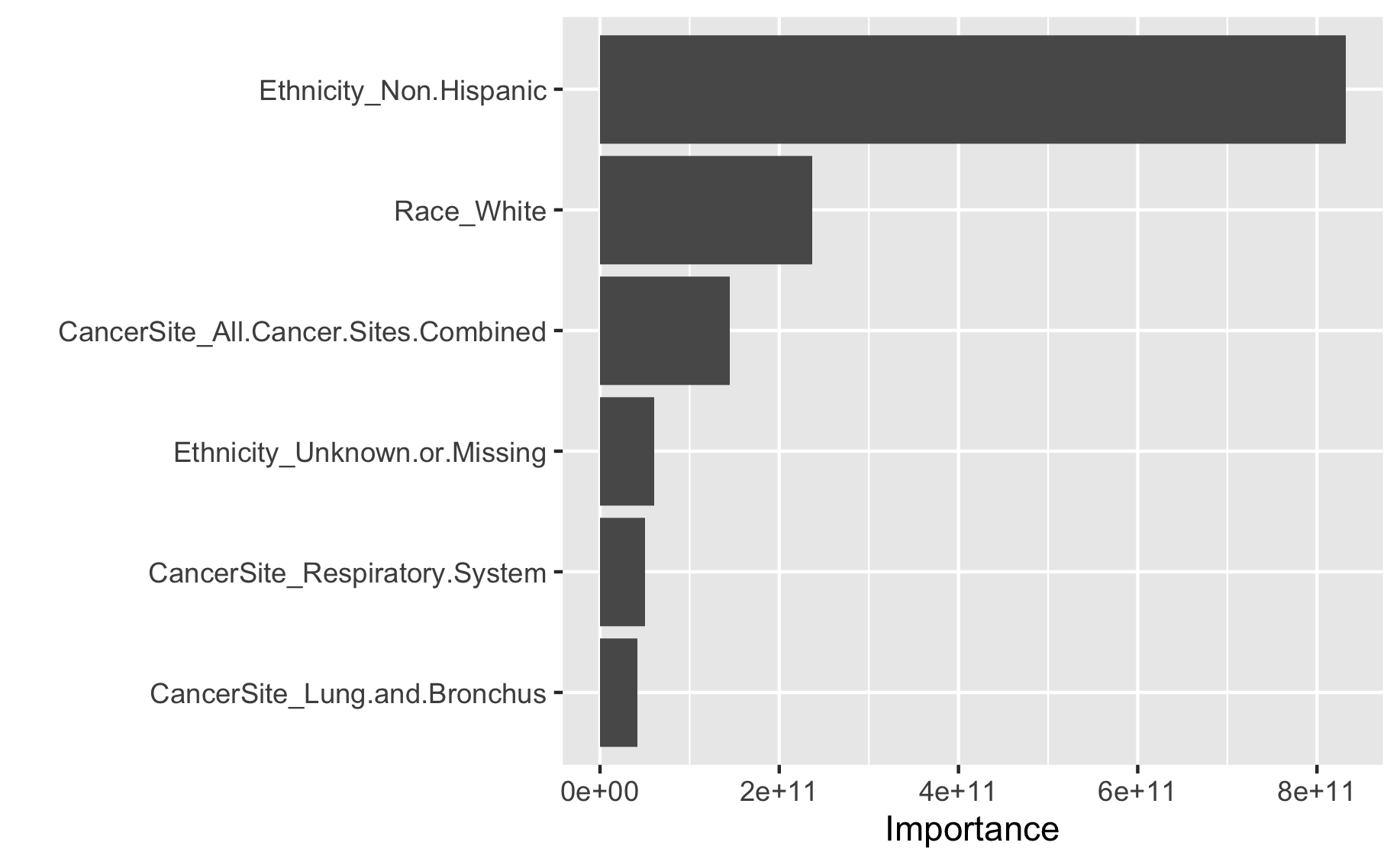


Figure 4.4: Variable Importance

### 4.0.1 Discussion

While the modeling overfit and was thus inconclusive that race and ethnicity play a role in cancer rates and death, the exploratory analysis and modeling suggests that there is a much higher prevalence of overall cancer death in the White population group, with the African American population group experiencing the second highest rates of cancer death overall.

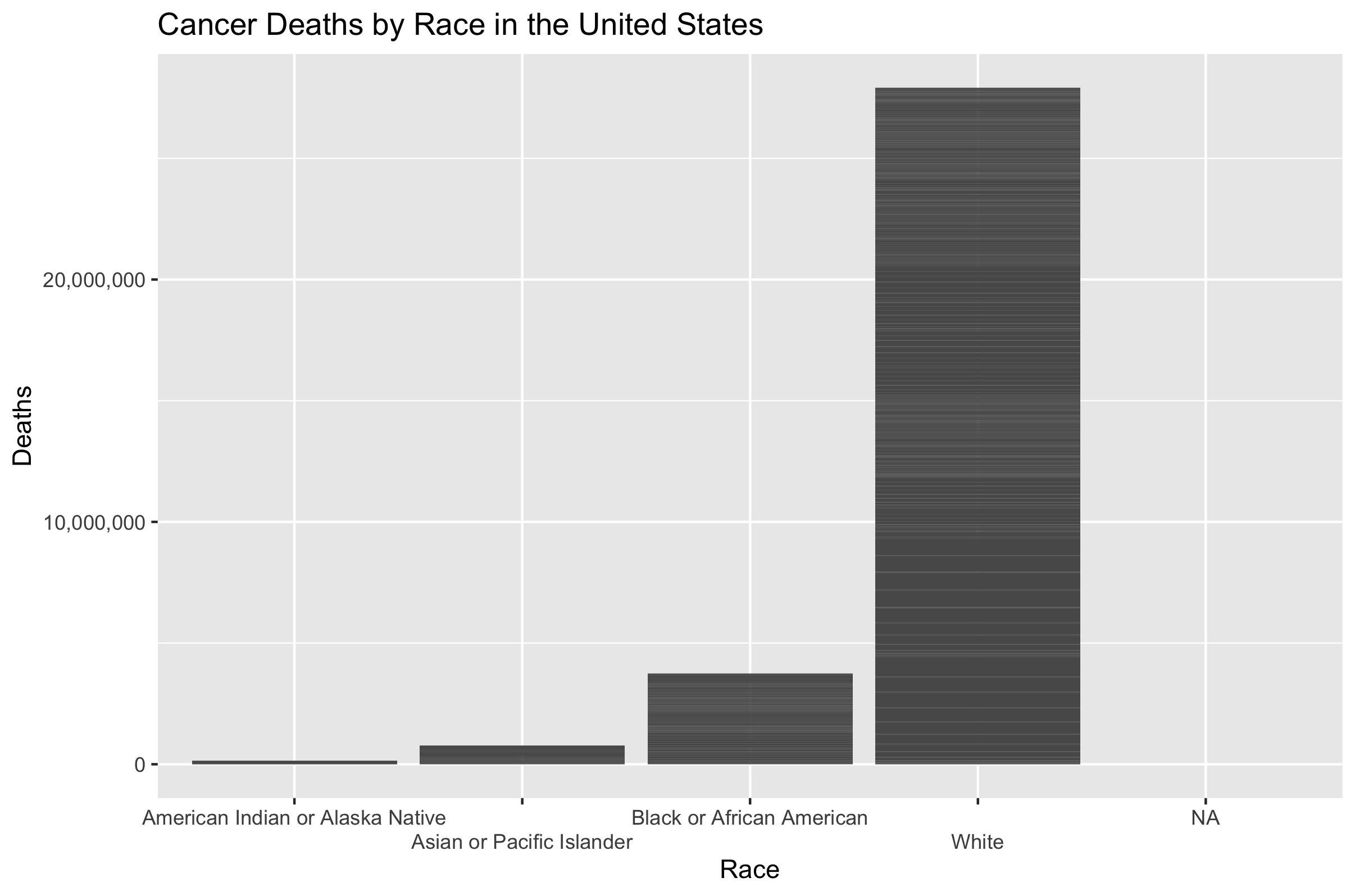


Figure 4.5: Cancer Deaths by Race

In terms of most common age group, age appears to have a direct link with an increased risk of developing cancer. In terms of sex, men appear to have more overall cases of cancer death than females.

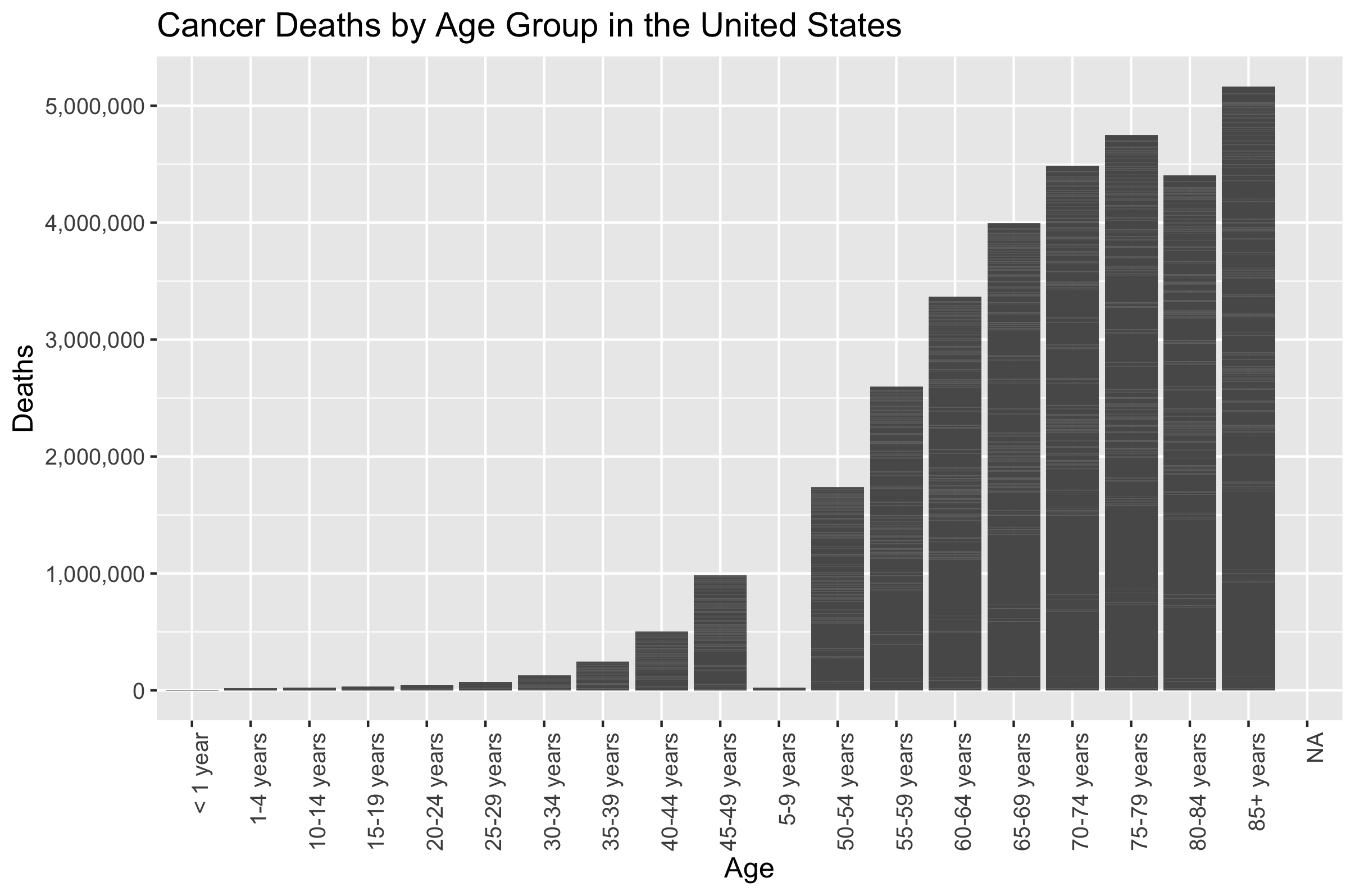


Figure 4.6: Cancer Deaths by Age Group

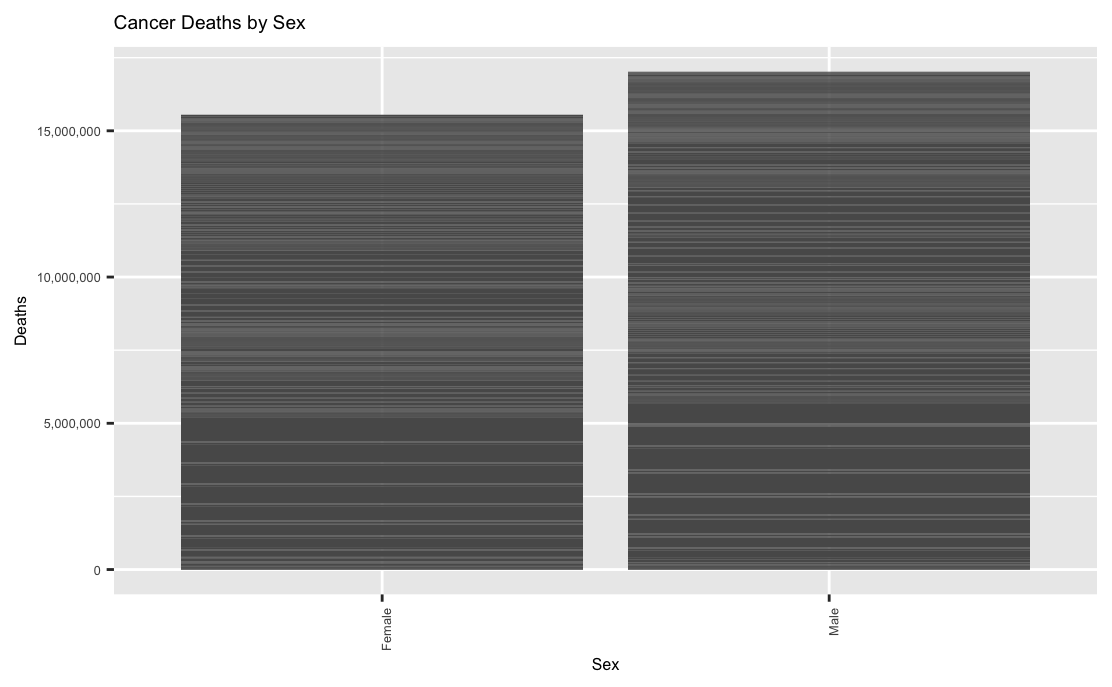


Figure 4.7: Cancer Deaths by Sex

Finally, when examining overall cancer sites, lung and bronchus folowed by the respiratory system were the two most common locations overall for developing cancer, with the digestive system being the third highest.

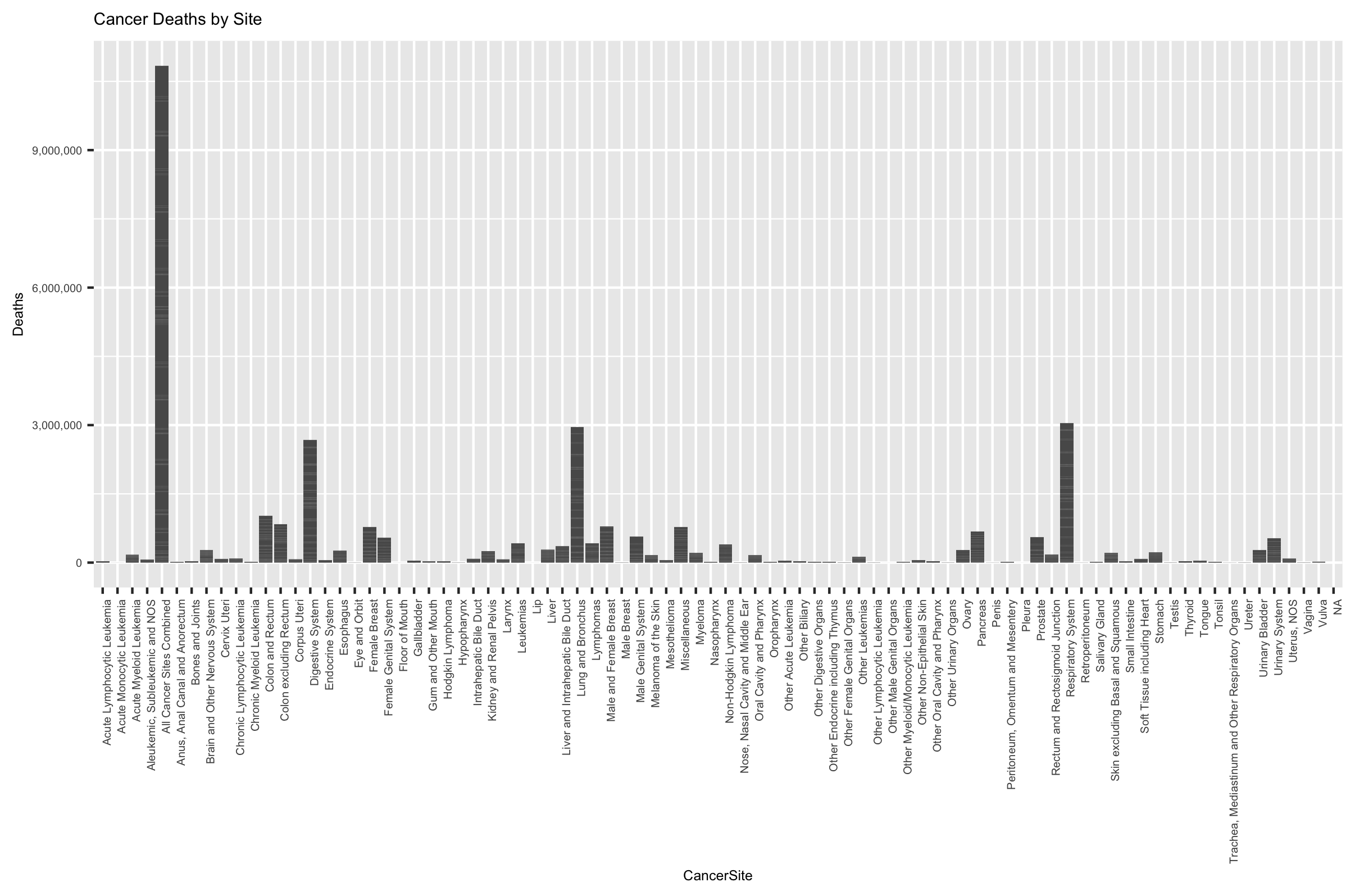


Figure 4.8: Cancer Deaths by Site

## 4.1 Summary and Interpretation

Overall, while no models were conclusive as to establishing a causal link between race, ethnicity, and cancer rates and deaths, there appears to be a distinct difference in overall disease burden between those of varying race, age groups, sex, and the location of where cancer appears. More research on race and ethnicity and how it relates to the burden of disease of certain population group should be pursued further to determine if any causal link exists.

In addition, care must be made to ensure an accurate portrayal of overall disease burden on various population groups is conducted. For instance, although there appears to be a much higher cancer disease burden in the White population, there may be a large difference in the reported numbers of cases due to differences in the majority population throughout the United States as opposed to the average global population racial and ethnic composition.

## 4.2 Strengths and Limitations

Strengths of the analysis are that the data comes from the Center for Disease Control, which enables the retrieval and usage of accurate and highly valuable public health data. Additionally, by focusing on the United States population in particular, possible trends relating to cancer death in the nationwide population can be examined and studied which may be hidden under a global lens. This unfortunately acts also as a limitation, in that the lack of data retrieved globally limits the overseas implication of any research results that can be extrapolated from the research project.

In addition, by relying on data already collected by the CDC rather than independently collecting data, there exists a possibility of certain population groups (for instance, those in rural areas) to be under represented overall.

## 4.3 Conclusions

In conclusion, while no clear evidence was proven in the modeling conducted in this project towards the relationship between race and ethnicity on cancer deaths, exploratory analysis indicates that both appear to have a tangible and measurable impact on the disease burden of a community, either directly through cultural stigma and institutionalized norms, or indirectly through socioeconomic status, family cohesion, and community networks. More research needs to be done to conclusively determine the presence of and extent of this potential relationship in the United States, or if the two predictors are simply confounders to another greater driver.

# References

[1] H.K. Weir, T.D. Thompson, A. Soman, B. Møller, S. Leadbetter, “The past, present, and future of cancer incidence in the united states: 1975 through 2020,” *Cancer*. 121 (2015) 1827–1837. <https://doi.org/10.1002/cncr.29258>.

[2] V.A. Zavala, P.M. Bracci, J.M. Carethers, L. Carvajal-Carmona, N.B. Coggins, M.R. Cruz-Correa, M. Davis, A.J. de Smith, J. Dutil, J.C. Figueiredo, R. Fox, K.D. Graves, S.L. Gomez, A. Llera, S.L. Neuhausen, L. Newman, T. Nguyen, J.R. Palmer, N.R. Palmer, E.J. Pérez-Stable, S. Piawah, E.J. Rodriquez, M.C. Sanabria-Salas, S.L. Schmit, S.J. Serrano-Gomez, M.C. Stern, J. Weitzel, J.J. Yang, J. Zabaleta, E. Ziv, L. Fejerman, “Cancer health disparities in racial/ethnic minorities in the united states,” *British Journal of Cancer*. 124 (2021) 315–332. <https://doi.org/10.1038/s41416-020-01038-6>.

[3] L.A. Torre, R.L. Siegel, E.M. Ward, A. Jemal, “Global cancer incidence and mortality rates and trends - an update,” *Cancer Epidemiology Biomarkers and Prevention*. 25 (2016) 16–27. <https://doi.org/10.1158/1055-9965.EPI-15-0578>.

[4] H. Dana, C. Greta, M. Matteo, B. Paola, W. Samuel, N. Eva, V.L. Carlo, B. Paolo, “Aging-12-103503,” (n.d.).

[5] C. Zhang, C. Zhang, Q. Wang, Z. Li, J. Lin, H. Wang, “Differences in stage of cancer at diagnosis, treatment, and survival by race and ethnicity among leading cancer types,” *JAMA Network Open*. 3 (2020). <https://doi.org/10.1001/jamanetworkopen.2020.2950>.

[6] M.B. Schabath, W.D. Cress, T. Muñoz-Antonia, Racial and ethnic differences in the epidemiology of lung cancer and the lung cancer genome, n.d.

[7] B.C. Özdemir, G.P. Dotto, “Racial differences in cancer susceptibility and survival: More than the color of the skin?” *Trends in Cancer*. 3 (2017) 181–197. <https://doi.org/10.1016/j.trecan.2017.02.002>.