Health Data Science 5310

Mortality Rate of Colorectal Cancer in Geriatrics Population Across United States

Anh Vu, Hao-Tu Yu, Victor Okoye

Saint Louis University School of Medicine

April 2023

Background:

Colorectal cancer is the third most commonly diagnosed cancer worldwide, with an estimated 1.8 million new cases and 881,000 mortalities in 2018 alone. [1] Disparities in colorectal cancer incidence among US states and outcomes can be due to numerous factors. At the individual level, demographic factors, occupation, poverty, and health behaviors or beliefs can impactfully contribute to cancer risk. [2] Additionally, there are many well-established cancer risk factors beyond the individual level, including healthcare access, living environment (social and physical), and large-scale policy and systems. [3] Despite advances in treatment and screening, the incidence and mortality rates of colorectal cancer remain high, making it a significant public health concern, especially within the African American communities. [4] Some factors that have been linked with early-onset CRC include obesity, physical inactivity, and unhealthy dietary patterns. According to data from a National Institutes of Health/AARP survey, each of these risk factors may be more prevalent in African American communities, particularly those of low socioeconomic status. [3] African Americans are also less likely to have access to health care. As a result, they may not be able to promptly seek medical care for symptoms associated with colorectal polyps or cancer. This, in turn, could delay detection of tumors, which could be cured by removal through colonoscopy or surgery if detected early enough. African Americans are less likely to receive chemotherapy or surgery compared to white patients. Currently, in the United States (US) residents in rural or non-metropolitan areas experience a greater cancer burden for tobacco-related cancers and cancers that can be prevented by screening. [2] While national trends show the rural–urban disparity in colorectal cancer (CRC) incidence has narrowed greatly over several decades, disparities remain. [4] The purpose of our research aims to investigate the relationship between certain risk factors and the development of colorectal cancer. Specifically,

the study will focus on the impact of age, sex, and ethnicity in the mortality rate of developing colorectal cancer.

While previous studies have examined the association between these risk factors and colorectal cancer, this research project will differ in its approach by examining additional variables such as age, gender, and ethnicity. By including these variables, the study aims to identify which demographics or populations are vulnerable, and provide a more comprehensive understanding of the factors that contribute to the development of colorectal cancer.

To collect the necessary data, this study will utilize a combination of electronic medical records from Real World Data from EMR of patients diagnosed with Colorectal cancer from 2013 - 2017. The dependent variable in this study will be the mortality rate of colorectal cancer, while the independent variables will be age, gender, and ethnicity. These independent variables were chosen based on their relevance to the development of colorectal cancer and their availability in the dataset.

The practical implication of this research is significant. If successful, we can present the data and inform policymakers to construct systematic and cost-efficient public health policies and clinical guidelines on the prevention of colorectal cancer. Additionally, the findings can be used as a tool to educate the general public about the risk factors in particular demographics and the importance of early detection for colorectal cancer.[5]

Methodology:

The objective of the analysis is to assess the disparity in the mortality rate of colorectal cancer among different ethnic groups in the United States. The independent variables are the mortality rates of colorectal cancers. To achieve this goal, Analysis of Variance (ANOVA) was

employed as the statistical technique. One-way ANOVA is an analytical method utilized for comparing means across three or more groups with similar categorical attributes. The ANOVA's F-statistic is a ratio that compares the variation between groups to the variation within groups. In order to conduct the analysis, we must formulate the null hypothesis and the alternative hypothesis:

H0 = the mortality rates of colorectal cancers are equal across ethnic groups.

HA = the mortality rates of colorectal cancers are not equal across ethnic groups.

The difference between the group mean and the overall mean represents the degree to which the group mean accurately reflects the data in that group. This difference determines the numerator of the F-statistic. The numerator of the F-statistic quantifies the variation between the group means and the grand mean, while the denominator quantifies the variation between individual values and the group means. The F-statistic can be expressed as:

F = (between-group variability) / (within-group variability)

A larger F-statistic indicates that the model has explained more of the variance in the data compared to what is left unexplained. In other words, a higher F-statistic implies that the difference between the means of the groups is more significant relative to the variation within the groups. The F-statistic can also represent the variance ratio among groups to those within the groups. A small p-value would indicate the F-statistic is statistically significant. However, an abundance of observations at the lower end of a range, known as a floor effect, or an abundance

of observations at the upper end of a range, referred to as a ceiling effect, can often affect the reading. As ANOVA is a procedure that assesses both mean and variation, the presence of floor or ceiling effects may mislead into finding a difference when there is not, or hinder the detection of differences even when they exist. This highlights the importance of recognizing and addressing such limitations in statistical analysis.

Before conducting an one-way ANOVA test, it is essential to run preliminary tests to ensure that these assumptions are met:

- Independence of observations: observations must exclusively belong to a single group, with no correlation or dependency between observations in each group.
- Independence of groups: groups must have no relation or dependency on one another.
- Normality: all outcomes should be normally distributed within each group.
- Homogeneity of variances: equal variance of the outcome across all groups.

The normality assumption can be checked by using either visually by using a histogram, density plot, or a Q-Q plot to identify a normal or non-normal distribution, or statistically with the Shapiro-Wilk test to examine the null hypothesis that the data are normally distributed. Levene's test is used to test the assumption of equal variances, with the null hypothesis being that the variances are equal, while the alternative hypothesis is that at least two of the variances are different.

When the normality assumption is met but the homogeneity of variances fails, we use the Brown-Forsythe or the Welch's approach to calculate F-statistics. The Brown-Forsythe method for calculating F-statistics involves transforming the continuous variable from its measured

values to values that represent the distance each observation is from the median of the variable. This transformation is used to address the issue of unequal variances between groups. In contrast, Welch's F-statistic uses weights in the calculation of the group means and overall mean. The oneway.test() command can compute Welch's F-statistic when the var.equal = FALSE. When the normal distribution assumption fails for ANOVA, the Kruskal-Wallis test is used by comparing the medians of the groups instead of comparing means

To measure effect size, ANOVA often employs eta-squared ($\eta 2$) and omega-squared ($\omega 2$). We will favor using the omega-squared to compute the effect size since it is adjusted to account for any positive bias and is more stable when assumptions are not met.

ANOVA is classified as an omnibus test, which indicates that a significant outcome implies the presence of a mean difference. However, it does not specify which means are statistically significantly different from another. In order to determine the statistical significance between means, ANOVA employs planned contrasts and post hoc tests. The Bonferroni post-hoc test is used to conduct a t-test for each set of means while also adjusting the threshold for statistical significance to maintain a low risk of Type I error, which is when the statistical analysis rejects the null hypothesis when the null hypothesis is true. Tukey’s Honestly Significant Difference (HSD) post-hoc test can also offer a less conservative insight. However, Bonferroni and Tukey's HSD post-hoc tests are not intended to categorize the groups and compare their respective means. This can be achieved using planned comparisons, which compare any subset of means to another subset of means. This enables us to perform more targeted comparisons and analyze specific significant differences between groups in a comprehensive manner.

To facilitate data interpretation, box plots were selected for data visualization as they offer a more comprehensive understanding of the distribution of data within each group and provide a better portrayal for a comparative analysis between the groups. Histogram, density plot, or Q-Q plot will be utilized to identify the nature of the distribution of each group of interest.

Linear regression is utilized to predict the continuous dependent variable based on a set of independent variables. In this case, the dependent variable is the mortality rate for colorectal cancer, and the independent variable is ethnicity. To perform linear regression analysis, several assumptions must be met, including:

- Linearity: The relationship between the dependent variable and independent variable(s) should be linear.
- Independence: Observations should be independent of each other.
- Homoscedasticity: The variance of the residuals should be constant across all levels of the independent variable(s).
- Normality: The residuals should be normally distributed.

If these assumptions are met, we can use linear regression to estimate the slope and intercept of the best-fit line that describes the relationship between ethnicity and the mortality rates of colorectal cancer. We can compute the slope coefficient to determine the magnitude and direction of the relationship between ethnicity and the mortality rates of colorectal cancer. Additionally, p-value and confidence interval (CI) can be used to assess the significance of the slope coefficient. Lastly, we can evaluate the overall fit of the linear regression model using

various measures, such as the coefficient of determination R-squared, which provides information about the proportion of variance in the dependent variable that is explained by the independent variable(s). This line can then be used to make predictions about the mortality rates of colorectal cancer for different ethnic groups.

The insights gained from this analysis can be used to redirect efforts towards exploring the risk factors associated with colorectal cancer and developing policies and interventions to support vulnerable communities. For example, if the analysis reveals that certain ethnic groups are at higher risk for colorectal cancer, policymakers and healthcare providers can focus on providing targeted screening, prevention, and treatment programs to those communities.

We also look at correlations between variables in the cancer data by sex and age, specifically mortality rates in men and women over 64 years of age, because of the increased risk of cancer in this age group. In this observation, we want to set gender as an independent variable to see if there are significant differences in cancer mortality rates by age. We will use an independent sample t-test to compare the means of the two groups with each other.

The following is a review and examination of the underlying assumptions for the use of t-tests:

- Continuous variable and two independent groups
- Independent observations
- Distributions must be normal
- Variances in each group must be equal

A histogram or a Q-Q plot can be used to visually distinguish between normal and abnormal data distribution. A Shapiro-Wilk test can be applied statistically.

We can test the assumption of homogeneity of variances by Levene's test. To verify the assumption of equal variances, Levene's test is frequently employed. The variances are equal, which is the null hypothesis for Levene's test; the variances are not equal, which is the alternative hypothesis. A statistically significant Levene's test would result in the rejection of the assumption and the null hypothesis of equal variances.

When the independent-samples t-test normality assumption fails: The Mann-Whitney U test

When the independent-samples t-test variance assumption fails, we can use the Kolmogorov-Smirnov test

To establish a relationship between age and gender with an increased mortality rate of colorectal cancer, an independent sample t-test is conducted to compare the distinct groups. The null hypothesis states that there is no significant difference between the mean age of female/male over the age of 64 and an increased mortality rate of colorectal cancer; whereas the alternative hypothesis states there is a significant difference between the mean age of female/male over the age of 64 and an increased mortality rate of colorectal cancer:

H0: There was no difference in the mean mortality rate between males and females at age 64 and older.

HA: There is a difference in the mean mortality rate between males and females over the age of 64.

When the number of variances of two populations are not equal, we can compare two independent population means by Welch's two-sample t-test, even in the case of unequal

variances. Then we compute the test statistics. In the independent sample t-test formula, m1 is the mean of one group and m2 is the mean of the other group; the difference between the means of the two groups forms the numerator. The larger the difference between the means of the two groups, the larger the numerator and the larger the t-statistic.When $p < .05$, we can interpret this to mean that if the null hypothesis is true, the probability of the value of this t-statistic occurring will be much less than 5%. Observing whether the difference in mortality rates between men and women over the of 64 is significant through p-values. We could find a significant difference between the mean mortality rates of males and females over 64 in the sample by the results [t-value; $p<0.05$]. What might be the value of the difference in mean mortality rate between males and females in the sample.

Results:

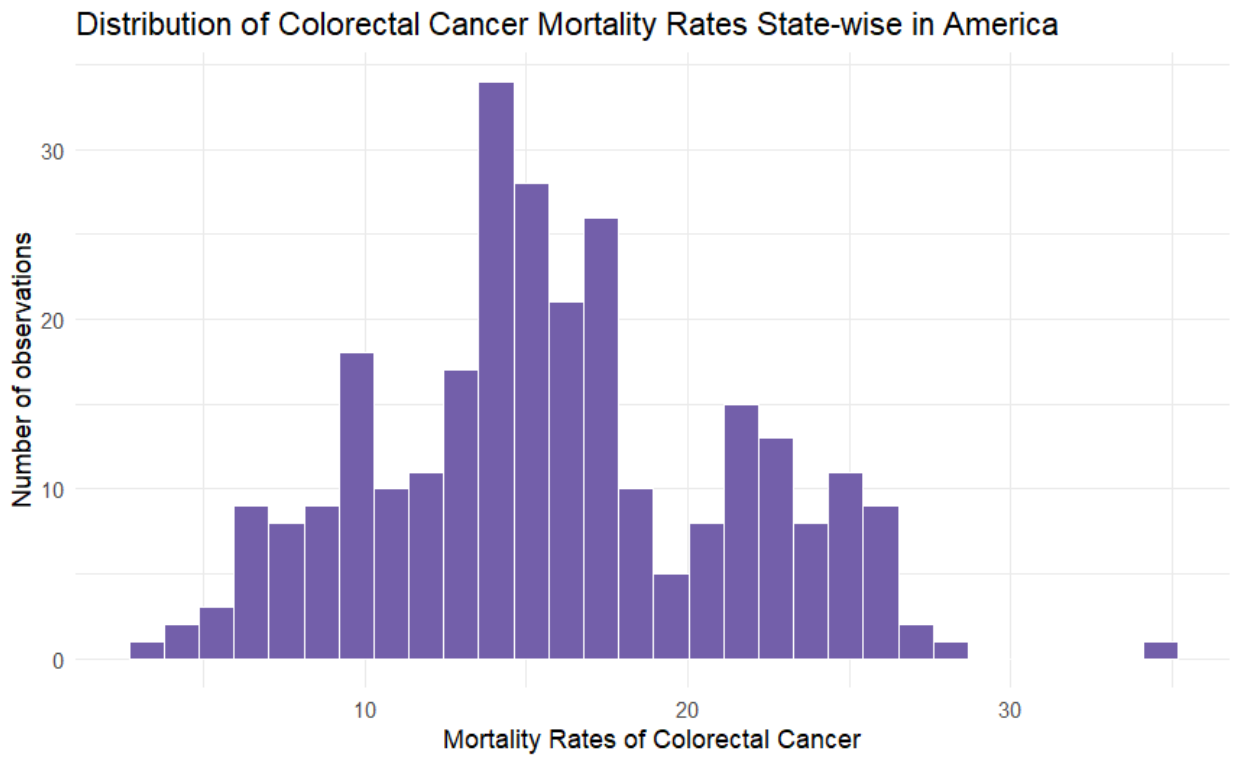

Figure 1. Distribution of Colorectal Cancer Mortality Rates State-wise in America.

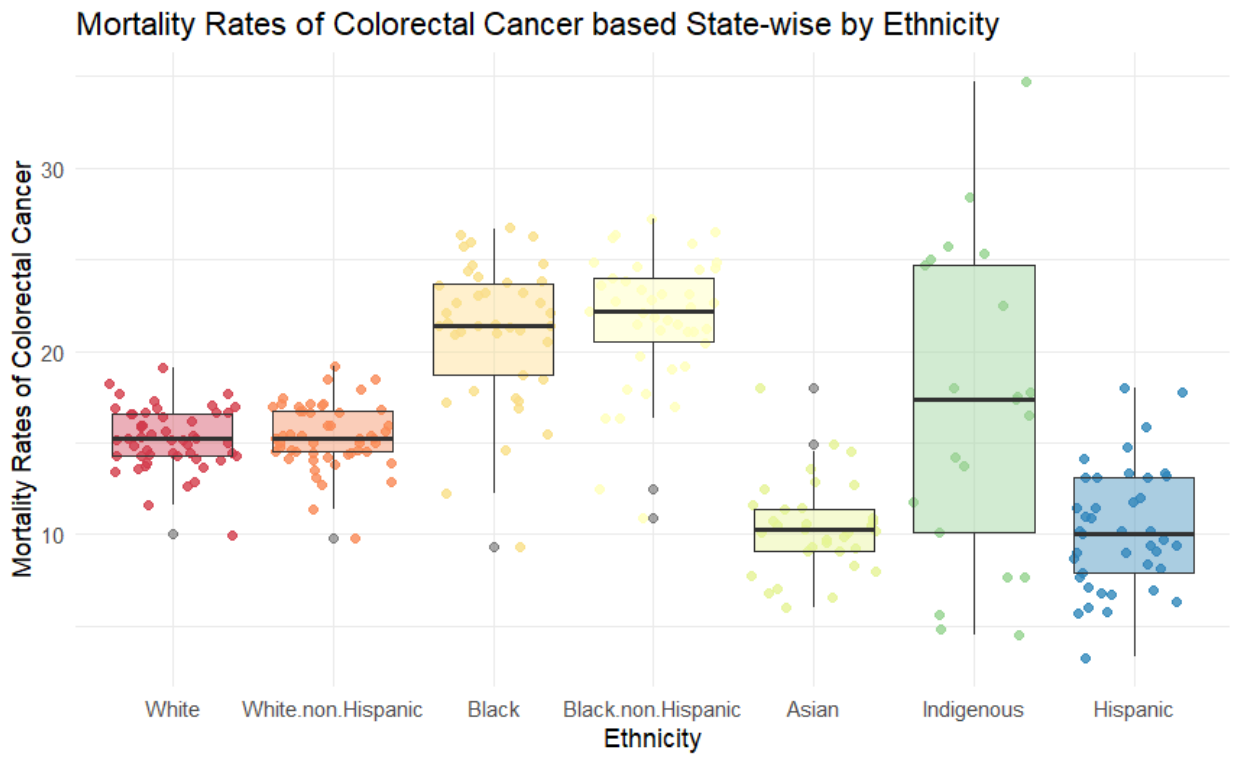

Figure 2. Assessment of State-by-State Incidence of Colorectal Cancer Deaths by Ethnicity. Bonferroni pairwise t-tests were used. Levene's test was used with unequal variances *$P < .05$. Kruskal-Wallis test is used instead **$P < .001$

Our study compared the average mortality rate of colorectal cancers in seven ethnic groups: White, White Non-Hispanic, Black, Black Non-Hispanic, Asian, Indigenous, and Hispanic. ANOVA one-way analysis of means indicated the average mortality rate of colorectal cancer is significantly different across ethnicity [$F(6, 273) = 63.653$; $p < .05$], indicating the disparity in vulnerability to cancer risk among the ethnic groups in America. The disparity of mean is most significant in Black (Mean = 21.16; SD = 3.86) and Black Non-Hispanic (Mean = 21.7; SD = 3.58) groups. Bonferroni testing showed there is a significant difference in average mortality rate of colorectal cancer between African American and non-Hispanic African American versus the rest of the groups ($p < .05$), between white and non-Hispanic white versus

the rest of the groups, except for the Indigenous ($p < .05$), and between Asian versus the Indigenous ($p < .05$).

Due to violations of the assumptions of equal variance and normality, alternate tests were used. The Welch test revealed that there is a significant difference in colorectal cancer mortality rates among the ethnic groups [$F_w(6, 102.42) = 72.307$]. The Kruskal-Wallis test found a statistically significant difference in mortality incidences of colorectal cancer across ethnicity groups ($H(6)=170.95$; $p<.05$). The relationship between ethnicity and colorectal cancer mortality rate is strong ($\omega 2 = .99$).

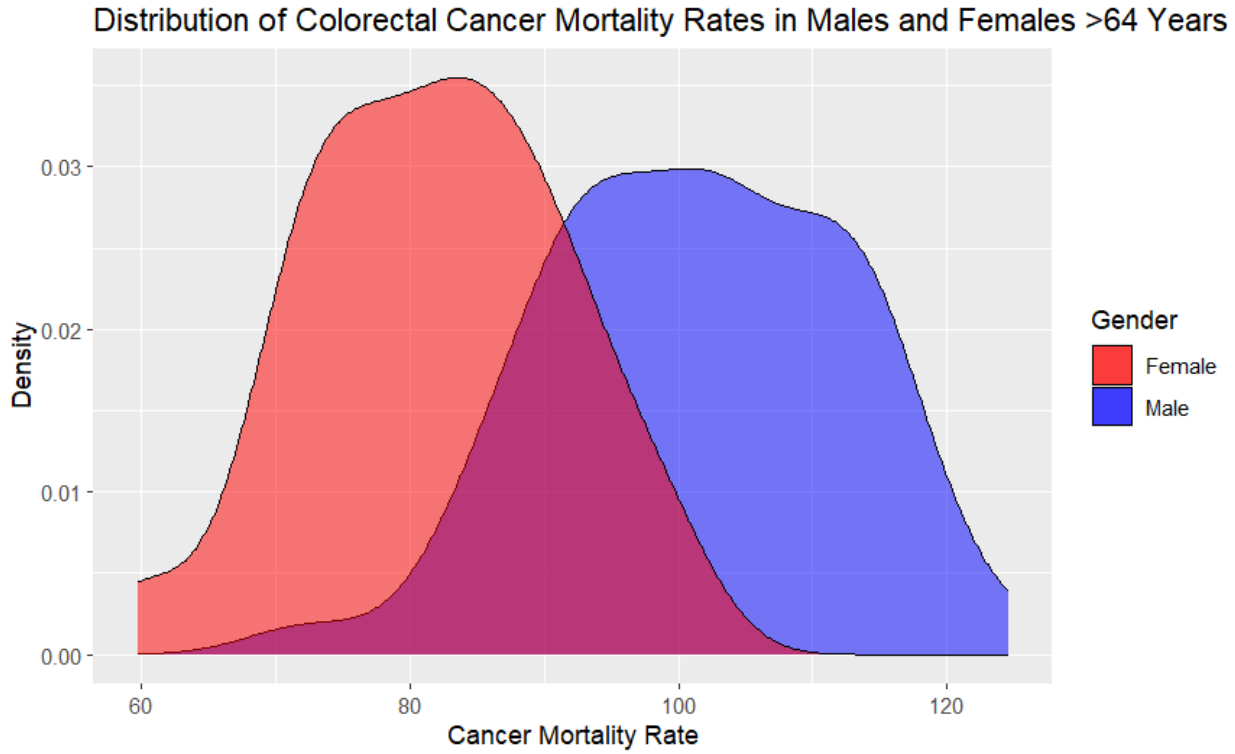

Figure 3. Density plot of colorectal cancer mortality rates in males and females over 64 [$t(98.26) = -9.7$, $^{***}P < 0.001$]. The F test and Shapiro-Wilk tests were used.

The density plot shows the distribution of colorectal cancer mortality rates in males and females over 64 years old using smoothed lines. We can see that the male distribution (in blue) is

skewed to the right compared to the female distribution (in red), indicating that, on average, males in this age group have higher mortality rates in colorectal cancer than females. The plot also shows the overlap between the two distributions, which highlights the range of mortality rate that are common to both males and females.

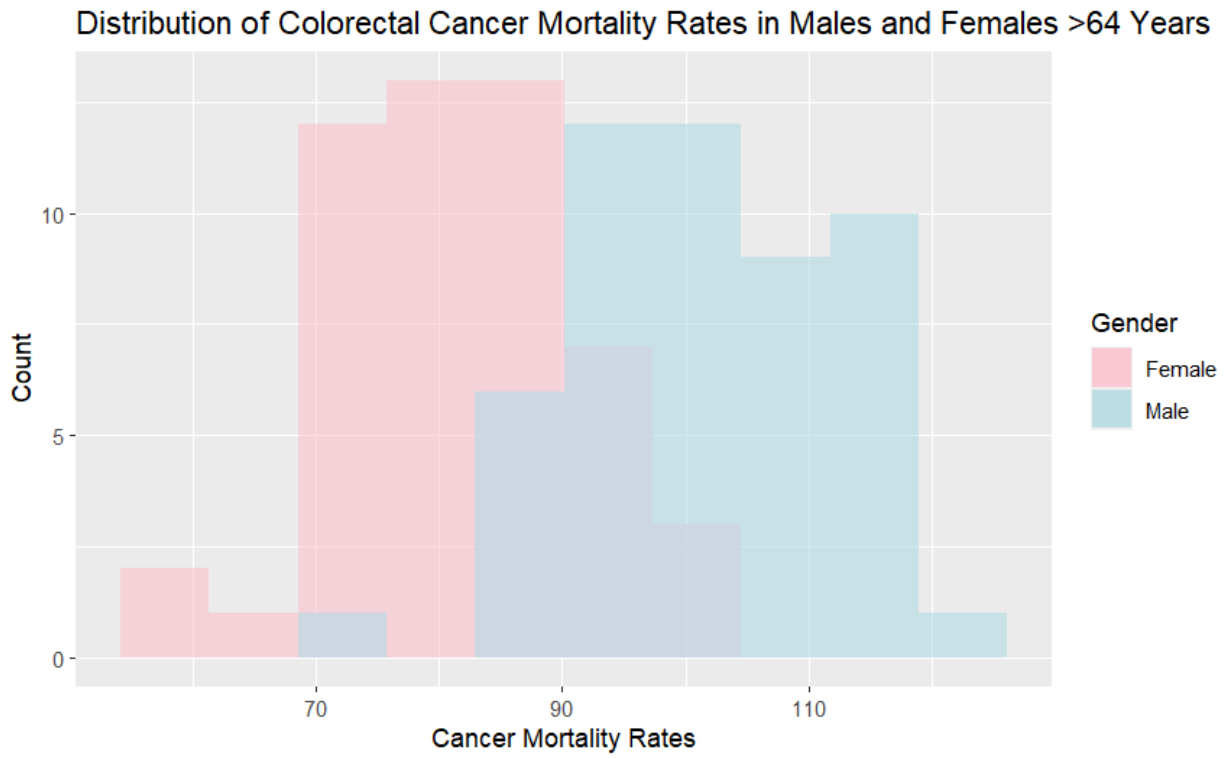

Figure 4. Histogram of colorectal cancer mortality rate in males and females over 64 [t(98.26) = -9.7, ***P < 0.001]. The F test and Shapiro-Wilk tests were used.

In Figure 3 and 4, the t-value shows how the means of the two groups differ in relation to the degree of data variability. The difference between the means of the two groups is bigger when the absolute t-value is larger. The t-value in this instance is -9.7. A warning symbol denotes that women are more likely than males to develop cancer on average.

The amount of evidence contradicting the null hypothesis, that there is no difference in the mean mortality rate between males and females at age 64 and older, is indicated by the p-value. Given that the null hypothesis is true, it shows the likelihood of observing a t-value that is equally extreme or more extreme than the one that was actually observed. Strong evidence is presented against the null hypothesis when the p-value is less than the significance level, which is often set at 0.05 [$t(98.26) = -9.7$, $p < 0.0001$]. The observed difference in mortality rate between males and females over the age of 64 is statistically significant in this instance since the p-value is zero. Since there is a significant difference in the mean mortality rate between males and females over the age of 64, the null hypothesis can be rejected. We concluded that there is strong evidence to support that the mortality of colorectal cancer in male over 64 is higher than that for their female counterparts.

Both the male and female data sets do not violate the normality assumption, according to the Shapiro-Wilk normality test results, given the p-values are higher than the significance level of 0.05. The variances of the two data sets are also not substantially different from one another according to the F test for equality of variances, as the p-value is higher than the significance level of 0.05.

The histogram of the female data reveals that the data has a peak at 100 and is slightly left-skewed. With a peak around 102, the male data histogram displays a more symmetrical distribution. However, the range of values shown by both histograms is comparable.

Both data sets' Q-Q plots reveal that the data points closely follow the 45-degree line, which suggests that the data are normally distributed. In overall, the distributions of the male and

female data sets are comparable, and they both meet the prerequisites for a two-sample t-test, which can be used to compare the means of the two data sets.

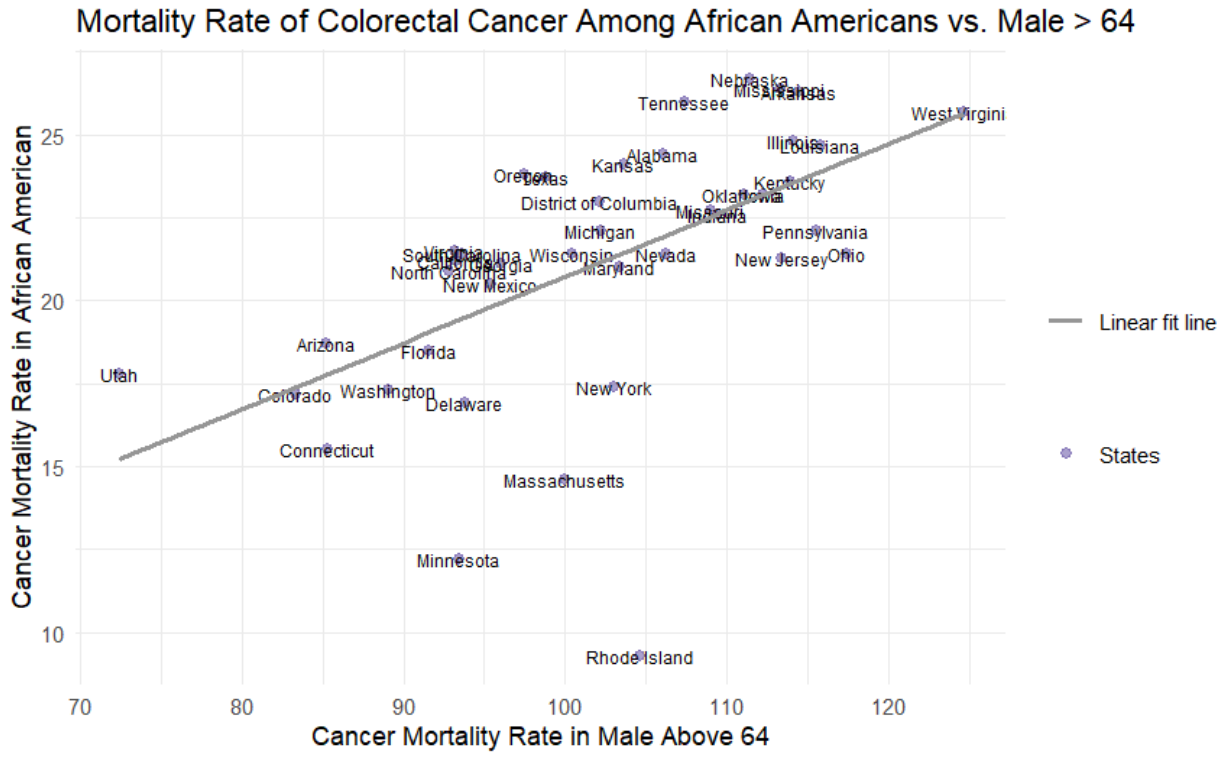

Figure 5. Assessment of Linear Regression of Colorectal Cancer Mortality Rates Between non-Hispanic African American vs. Male Above 64. Correlation coefficient was positive and moderately strong (r = 0.575).

Figure 5 highlights that the colorectal cancer mortality rates among African Americans in a state are a statistically significant predictor for males aged 64 and above ($b = 0.20041$; $p < .05$). For every 1% increase in mortality rates of African Americans in a state, the predicted mortality rate for male above 64 increases by 0.20041 percentage. The value of the slope in the sample is 0.20041, and the value of the slope is likely between 0.11 and 0.29 in the population that the sample came from (95% CI: 0.11–0.29). With every 1% increase in colorectal cancer mortality rate of African American, the mortality rate for male above 64 increases between 0.11 and 0.29 per 100,000 population. The correlation coefficient was positive ($r = 0.575$). These

results suggest that states with an older population and with a higher African American population experience a significant impact on their overall colorectal cancer risk, and more resources should be diverted to by the federal government into the funding of healthcare for those states.

Discussion and Conclusion

This research highlighted the demographic and ethnic mortality rates in colorectal cancer among different age groups from 2007 till 2013, revealing significant racial and ethnic disparities in the mortality of colorectal cancer in the African American communities in comparison to other ethnic groups. African American and African American Non-Hispanic individuals are 23-25% more likely to develop colorectal cancer. Among this population, males aged 64 and above have a significantly higher risk of all-cause mortality and colorectal cancer-specific mortality compared to their female counterparts. Although Indigenous groups had a significantly lower mortality rate of all-cause mortality and colorectal cancer-specific mortality than other ethnic groups, they have a higher variability in cancer mortality rate in comparison to other groups, possibly due to the erroneous practice of grouping all the different native American groups into one category. Also, we discovered a significant number of incidents occurring in Southern and Midwestern states that exceeded the linear fitness line. In contrast, East Coast and West Coast states experienced comparatively fewer incidents, with some exceptions. Our findings strongly suggest that patients of African American descent, particularly males over the age of 64, should undergo regular colorectal cancer screening. Early detection and treatment of colorectal cancer is associated with improved outcomes, including increased survival rates and a higher likelihood of successful treatment.

We have identified a number of deficiencies in the mortality reporting process across several states. Specifically, there are significant gaps in the reporting of patient outcomes data for African American, Asian, and Hispanic patients, as well as numerous instances of missing or zero values. By doing so, the true extent of disparities in cancer risk among different populations may not be adequately captured, leading to an underestimation or overestimation of cancer risk in certain states. For future studies, better reporting will facilitate interpreting and comparing results with other studies and help identify limitations of the analysis.

Additional research is necessary to pinpoint other factors that contribute to colorectal cancer disparities down in the county level, and to gain a deeper understanding of the underlying causes that shape these disparities. Colorectal cancer is a major public health concern in the United States, with significant disparities in its mortality and mortality rates among different racial and ethnic groups. Ourresearch has shown African Americans are more likely to develop and die from colorectal cancer than any other racial or ethnic group in the United States. Therefore, healthcare providers and policymakers should prioritize efforts to ensure communities at risk receive appropriate and timely colorectal cancer screening, which may include education and outreach initiatives, targeted screening programs, and increased access to screening services. By addressing these disparities, we can work towards reducing the burden of colorectal cancer in the African American community and improving overall health outcomes. The ultimate goal is to minimize, and ultimately eliminate, health disparities. A prospective cohort study could provide valuable insights by quantifying the potential improvements in population incidence and quality of life resulting from early colorectal cancer screening for all ethnic groups at risk.

References

1. Baidoun, F., Elshiwy, K., Elkeraie, Y., Merjaneh, Z., Khoudari, G., Sarmini, M. T., Gad, M., Al-Husseini, M., & Saad, A. (2021). Colorectal Cancer Epidemiology: Recent Trends and Impact on Outcomes. *Current drug targets*, *22*(9), 998–1009. https://doi.org/10.2174/1389450121999201117115717
2. Henley, S. J., & Jemal, A. (2018). Rural Cancer Control: Bridging the Chasm in Geographic Health Inequity. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, *27*(11), 1248–1251. https://doi.org/10.1158/1055-9965.EPI-18-0807
3. Zahnd, W. E., McLafferty, S. L., & Eberth, J. M. (2019). Multilevel analysis in rural cancer control: A conceptual framework and methodological implications. *Preventive medicine*, *129S*, 105835. https://doi.org/10.1016/j.ypmed.2019.105835.
4. Liang, P. S., Mayer, J. D., Wakefield, J., & Ko, C. W. (2017). Temporal Trends in Geographic and Sociodemographic Disparities in Colorectal Cancer Among Medicare Patients, 1973-2010. *The Journal of rural health : official journal of the American Rural Health Association and the National Rural Health Care Association*, *33*(4), 361–370. https://doi.org/10.1111/jrh.12209
5. Weitz, J., Koch, M., Debus, J., Höhler, T., Galle, P. R., & Büchler, M. W. (2005). Colorectal cancer. *Lancet (London, England)*, *365*(9454), 153–165. https://doi.org/10.1016/S0140-6736(05)17706-X

Data

1. Kafura, Dennis. (2019, June). Cancer CSV File [Type of report]. CORGIS Datasets Project. https://corgis-edu.github.io/corgis/csv/cancer/