**Race Matters: Analyzing the relationship between colorectal cancer mortality rates and various factors within respective racial groups**

Israel Almodóvar, Iowa State University

Monica Jackson, American University

Jada Johnson, University of Maryland

Jessica Lyle, Maryville College

Emma Veach, Michigan Technological University

Ismael Xique, Bowdoin College

**INTRODUCTION**

Colorectal cancer (CRC) is one of the most prevalent types of cancer in the United States. There are two forms of CRC, namely, one that attacks the colon or one that attacks the rectum. The American Cancer Society has estimated that there will be a projected 103,170 new cases of colon cancer and 40,290 new cases of rectal cancer in 2012. The chances of developing these cancers are about 1:20, however the risk is slightly lower for women than it is for men.9 CRC has one of the highest death rates among cancers. It is ranked as the third leading cause of death when men and women are considered as separate populations, and the second leading cause of death when both genders are combined.9 The average survival rate once the cancer has been discovered is approximately five years; most people, however, live much longer than five years, and many are often cured of the disease.9 The death rate has continuously decreased over the last twenty years for both men and women.9 This could be caused by more accurate and accessible screening, which results in earlier cancer detection and thus a higher rate of patient survival. Further, treatment of CRC has also been greatly improved over the years and, from these combined improvements, there are now over one million colon cancer survivors.9

While there are no definite causes of CRC, there are risk factors associated with the diagnosis of and mortality rates of CRC, including age, race, obesity level, and insulin concentration. Age is a key risk factor in the occurrence of CRC. Even though it can occur at young ages, the probability of developing colorectal cancer rises as a person reaches the age of 50 and it continues to grow as age increases.9 Another factor that heavily influences the risk of mortality from colorectal cancer is race. In the United States, the distribution of races varies throughout the nation. Whites disproportionally reside in the West and Northeast, while African Americans disproportionally live in the Southeast. It has been shown that residents in the South are less likely to get effective healthcare, which could justify the higher incidence rates of CRC in African Americans.1

A person’s obesity level is another trait that associates with his/her survival from CRC. Using body mass index (BMI) as a sole measure of obesity, the risk for CRC is higher among obese men than obese women. More generally, among men with elevated BMIs, CRC risk is positively associated with education level.6 For women, however, BMI is positively associated with the risk for mortality from CRC.6 Meanwhile, instead of using BMI alone, research suggests that it is more useful to measure the waist circumference along with BMI, due to the added correlation between the fat around the abdomen and the risk for colorectal cancer.3 CRC is also influenced by the concentration of insulin in the body. Insulin is one of the main causes for C peptide, which was found to raise the cancer risk threefold in women and even moreso in men.3 Studies have found that individuals with Type 2 diabetes have a higher risk of developing CRC.9 A meta-analysis looking at 20 studies investigating this association found a statistically significantly positive relationship between the two.11

There are other environmental factors that add to the risk of developing CRC. One of the main protectors against CRC is vitamin D. Americans lack a diet that is plentiful with vitamin D, so researchers are looking at the sun as a source for vitamin D.4 The area in which a person lives can affect the amount of vitamin D that the body is receiving from sunlight. Previous studies have shown that large metropolitan areas receive less sunlight due to the shadowing of buildings, whereas places that are more rural receive more direct sunlight. This direct sunlight possibly explains why people in rural areas appear to have lower risks and occurrences of CRC.4 Other studies consider the connection between profession and the amount of sunlight received. Indoor workers (e.g. doctors, lawyers, accountants) do not receive the same amount of sunlight as outdoor workers (e.g. farmers, cattle ranch owners, construction workers).4 The amount of sunlight exposure is negatively correlated with the risk of mortality from CRC.2

Social factors also influence the screening and mortality rates for colorectal cancer; one key contributor is education. A Tennessee study showed that people with at least a high school education are 2.47 times more likely to get screened for colon cancer than those without it.7 Meanwhile, annual screenings are shown to reduce CRC mortality by 33%, and biennial screenings reduce mortality by 21%.5 Thus, the fact that lack of education reduces the likelihood of screening infers that lack of education could also increase the chances of mortality. Another key social factor is a person’s residential area in the United States. Associations have been found between poor health and living in a county containing no cities larger than 10,000 people.8 While this link has not been made to CRC specifically, it could be a potential factor involved in increasing mortality rates from colorectal cancer. Another social factor that influences mortality rate is socioeconomic status. Some of the variation that exists between African-Americans and Caucasians can be accounted for by socioeconomic status.10

This paper is to further the knowledge of researchers with the spatial study, and to educate those who could be and are affected by the risk factors that could lead to colorectal cancer. The regression analysis will help us distinguish which traits should be emphasized as major contributors to risk of mortality. The results show differences between the races within certain risk factors.

**METHODS**

The statistical software package, *R*, was used for all analysis (*R* Development Core Team, 2012). R: A language and environment for statistical computing. *R* Foundation for Statistical Computing,Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.)

**Data**

The data for this study were obtained from three sources: the United States Census Bureau, the Centers for Disease Control and Prevention (CDC), and the National Solar Radiation Database; Table 1 lists the variable names of interest and their descriptions. Alaska and Hawaii were omitted from this study, because these states are isolated (i.e., they do not border any other states) and thus have no spatial correlations. Further, Broomfield County, Colorado (which is composed of parts of neighboring counties) and Clifton Forge City, Virginia (which is no longer recognized as a county) are also omitted for various reasons. Further, the natural-log transformation of the mortality rates for each of the races considered was used in order to approximate a normal response variable.

The Census Bureau provided information for the year 2000 regarding education level, population, socioeconomic status, income, population type (i.e. rural versus urban population), and poverty level by county. This data was used because there is an estimated five-year survival rate for colorectal cancer (ACS, 2012), thus more accurately representing the living conditions of people who died in 2005.

Education level, initially defined by three classifications (high school degree, bachelor’s degree, and graduate or professional degree), was condensed into one category (at least a high school degree or otherwise) to improve the efficiency of our model. The following CDC data from 2005-2007 was collected and averaged for each of the study years considered: percent of the population with diabetes ; percent of the population considered obese ; and the mortality rates of the Caucasian (non-Hispanic/Latino), Black or African-American, and Hispanic or Latino populations of colorectal cancer by county. The population data was limited to people 50 years or older because this age group is more at risk for CRC.

The National Solar Radiation Database (NSRDB) supplied the extraterrestrial radiation horizontal (ETRH) averages by state for 1991-2005. ETRH denotes the sunlight collected from a plate that is horizontal to the earth’s surface (3Tier). This data was used for this study because humans are more likely to identify with the horizontal measurement since they absorb the sun’s rays based on the angle of the sun and the time of day. Since we are considering detection and mortality rates from 2005-2007, we chose to analyze the most relevant (i.e., 2005) data (NSRDB, Site-File). The NSRDB collected data from Class I, Class II, and Class III airports. Class I airport data was used since it had the most extensive coverage of solar radiation (NSRDB, Manual). There were 221 Class I airports including Hawaii and Alaska. The airports in Hawaii and Alaska were omitted because these states are not being used in the analysis. Oklahoma does not have Class I airports, thus two Class II airports were chosen and averaged to calculate the temperature data for the state of Oklahoma.

The solar radiation data was composed of specific daily measurements (from 7:00 a.m. to 4:00 p.m), which were averaged for the entire month. The average of every month for each Class I airport was calculated, along with the yearly average for each airport. The yearly averages for the individual airports were used to calculate state radiation averages. Every Class I airport in a specific state was included in the calculation for a state average. County data for solar radiation could not be calculated because not every county had an airport, so the state average was alternatively used.

The CDC provided data on the county level mortality rates for all 3,143 counties in the 50 states and the District of Columbia. Gender and race were associated with mortality rates due to colorectal cancer from 2005-2007. The mortality file provided the national, state, and county-level population estimates based on the Census.

**Spatial Analysis**

A modified version of Moran’s I was used to test for geographic clustering for the three dependent variables. Modified Moran’s I is defined as

where and are the colorectal death rates at geographic locations and , respectively; is the expected colorectal death rate using all the data; and is the total number of geographic units. The weight function used is known as the population density adjusted exponential weight function (Thompson, 2011) and defined as

,

where is the Euclidian distance between the geographical points and , and can be set to small or large values and correspondingly “increases the sensitivity of the test to large or small clusters” (Thompson, 2011). We let to allow for possible detection of any sized clusters. The test statistic and corresponding p-value provide inference regarding the hypothesis test for statistically significant spatial correlation. This test was performed (via the *R* statistical analysis software) for each of the sample populations.

Spatial regression analyses were used to examine the relationship between colon cancer mortality rates and the geographical, weather, and socio-economic variables presented in Table 1 for the spatially correlated data. Linear regression was meanwhile performed to analyze the nonspatially correlated response data. Separate regression models were built for each the three dependent outcome variables related to the mortality rates by county for Hispanic, African-American, and Caucasian groups, respectively. Table 2 provides the summary statistics for each of the independent and dependent variables used in the regression model. We fit a simultaneous autoregressive model which uses a regression on the values from all locations based on an associated weight function to account for the spatial dependence to estimate the variables that were predictors for colon cancer mortality in each county.

Table 3 presents the Pearson’s correlation coefficients for all pairs of variables considered in the analysis. The strong correlation between some of the variables was expected (e.g., male and female genders; and obesity and diabetes). Due to the high correlation between male and female genders, we choose to focus our attention on males since colon cancer is more of a risk to men. Given the high correlation between obesity and diabetes, we only considered obesity in our model, since obesity is a curable factor and can possibly be altered in those affected.

**RESULTS**

**Descriptive Statistics**

This study examined the relationship between colon cancer morality death rates and several variables regarding atmosphere, geographic location, and socioeconomoic status. Descriptive statistics reveal that the approximate mean death rates across all counties are 7.5% for the African-Americans, 1.17% for Hispanics, and 1.61% for Caucasians, respectively; see Table 2. The largest mortality rates for each of the respective races occurs in Santa Cruz County, Arizona (for African-Americans); Douglas County, South Dakota (for Hispanics); and Grayson County, Kentucky (for Caucasians).

**Spatial Regression**

Modified Moran’s I was used to examine the spatial relationship between neighboring regions. Using the adjacent neighbor matrix defined in Equation (1), we obtained values 0.008 (p-value = INSERT) for African-Americans, 0.002 for Hispanics, and 0.021 for Caucasians, respectively. These results imply that African-Americans and Caucasians showed clear spatial relationships, while such a relationship was not statistically significant among Hispanics.

A separate spatial analysis was done for the colorectal cancer rates for each race considered in this study. After a careful analysis of the explanatory variables (Table 2) and their correlation table (Table 3), we choose to use average sunlight, percent of obesity, percent rural of the county, prevent below poverty, and male over fifty in the analyses. These variables were chosen in light of the strong correlations evident in the correlation table for some of the variables; see Table 3. Obesity and diabetes were strongly correlated, as well as percent white over fifty with the percent of high school degrees or higher. Urban and rural, as well as male and female were highly correlated. These correlations are easily explained due to the fact that they are complementary factors. Thus, we consider the following spatial model,

to associate the various explanatory variables (ETRH, percent of the county deemed rural, obesity rates in the population, and percent of population with at least a high school degree, percent below poverty level, and percent of the population that is male and over 50 years old) and spatial relationships with the CRC mortality rate within each respective race.

The data for African-Americans showed significant spatial associations. A spatial linear regression model was then calculated using data that connected neighboring counties to each other. The model was statistically significant (test statistic equals 0.1989; p-value < 0.001), implying that the spatial model is informative. In particular, four traits (ETRH, percent of the county deemed rural, obesity rates in the population, and percent of population with at least a high school degree) were found to be statistically significant in the model concerning the African-American mortality rate; see Table 4. The CRC mortality rate among African-Americans was found to be positively correlated with ETRH (p-value < .001) and average percent of the population that is obese (p-value = .0019), and negatively associated with the percent of the county that is classified as rural (p-value < .001) and percent of population with a high school degree (p-value = .0188).

No statistically significant spatial associations were detected within the Hispanic population. Statistically significant spatial connections were, however, detected among Caucasians; see Table 5. The model was statistically significant (test statistic equals 0.16485; p-value < 0.001), inferring that the spatial model is informative. In particular, five of the six traits considered (i.e., the percent of the population falling below the poverty level, average solar radiation, average obesity rate, percent of the county that is rural, and the percent of the population that is male) are statistically significant. The CRC mortality rate among Caucasians shows a positive association with the percent of the population that fell below poverty level (p-value = 0.044), average obesity (p-value < 0.001), and the percent of the population that was male (p-value <.001). Meanwhile, the CRC mortality rate is negatively associated with average solar radiation (p-value = 0.001) and the percent of the county that is rural (p-value < 0.001).

**Multiple regression**

The modified Moran’s I did not indicate the existence of spatial clustering of CRC mortality rates among Hispanics, therefore we consider a multiple regression model, namely

to describe the relationship between CRC mortality rates among Hispanics and the various explanatory variables considered (ETRH, percent of the county deemed rural, obesity rates in the population, and percent of population with at least a high school degree, percent below poverty level, and percent of the population that is male and over 50 years old). Four of the six traits analyzed in the model (namely, the percent of men over fifty, the percent of the county classified as rural, average obesity level, and the average amount of sunlight) were statistically significant; see Table 6. More precisely, the CRC mortality rate among Hispanics was positively correlated with percent of men over fifty (p-value = .0013), average obesity level (p-value = .0095), the average amount of sunlight (p-value = .0331), and the percent of the county classified as rural (p-value = .0081). The resulting regression equation is thus

The corresponding residual analysis showed some slight discrepancies, but primarily showed relatively normal residuals, reasonably satisfying model assumptions.

**DISCUSSION**

This study is significant because it not only identifies factors associated with the risk of colorectal cancer mortality but, more importantly, demonstrates how these factors vary within different racial groups. Accordingly, education on reducing risk factors for CRC should be directed at specific racial groups above and beyond creating a generalized education plan.

Obesity is consistently recognized as having a statistically significant positive association with CRC mortality rates across all racial groups. While there is already a demonstrated need to address the increase of obesity in the United States, the suggestive link between obesity and colorectal cancer provides yet another significant reason to promote better health in the United States. The percent of respective counties considered rural is another significant factor across all races. Though obesity was a significant factor across all racial groups, the direction in association with CRC mortality varied by race. Among African-Americans and Caucasians, the association was negative, while there was a positive association with the CRC mortality rate in the Hispanic population.

Sunlight also played an interesting role in this study. It was found to be statistically significant in models, increasing the CRC mortality rate among Hispanics and African-Americans, but decreasing the rate among Caucasians. One possible explanation for this difference is skin color. Individuals with darker skin tones absorb more sunlight, which can be harmful and cause other cancers and health issues rather than provide the beneficial effect of vitamin D absorption. It is also known that a higher percentage of African-Americans live in the southern areas of the United States. Since sunlight is significant in the spatial model, it is possible that (by living in the south) African-Americans receive sunlight amounts that surpass vitamin D requirements. Caucasians living in the south, however, may require more sunlight exposure in order to absorb the appropriate amount of vitamin D. Another potential explanation for the difference across races is their diets. Hispanics and African-Americans may have higher levels of vitamin D in their diets, so the beneficial effects of sun absorption may not be applicable in decreasing the CRC mortality rate. Meanwhile, Caucasians may have diets lower in vitamin D so the sunlight is needed to make up for the dietary discrepancy in vitamin D and help to act against colorectal cancer.

Some of the results obtained contradict the results of previous studies. One study states that sunlight decreases the risk of CRC mortality, however, this study finds a positive association between sunlight and CRC mortality risk in minority populations (which, as stated earlier, could be due to diet or skin color. Another study found that living in a rural area decreases the amount of access you have to health care and thus increases mortality rates. Our two spatial studies, however, detect a negative association between living in a rural area and CRC mortality rates. This could be because of the different living habits of those residing in rural areas. Urban areas are often faster paced and have different diets and behavioral patterns, which could negatively affect the individuals living there. Furthermore, cities have more shade coverage (particularly from tall buildings) and people living in cities are thought to be outside less often than those living in rural places. Living in rural areas decreases the risk of mortality from colorectal cancer in both Caucasian and African-American populations, and increases the mortality rates among Hispanics.

The results obtained associating obesity to CRC mortality support previous research. In this study, education was a statistically significant factor only among African-Americans; similar conclusions were obtained in INSERT. The relationship between education and other races has been unsubstantiated. According to the American Cancer Society, males are at an increased risk of both CRC incidence and mortality; this study confirmed these results among Hispanics and Caucasians. The percent of individuals below the poverty line caused an increase in CRC mortality rate, but only among Caucasians. This supports the previous research done on the relationship between socioeconomic status and mortality rate. Possible explanations for this relationship include patient willingness to seek medical attention, and medical facilities required to provide medical attention to emergency patients, whether or not the patient is insured. This, however, does not explain why poverty is only a significant factor among Caucasians.

**Limitations**

One of the major limitations of the data was the amount of zeros contained in the African-American and Hispanic datasets. Finding an appropriate transformation that would create more normalized data was difficult -- only an appropriate transformation for the Caucasian dataset was found. Accordingly, for the African-American and Hispanic datasets, we also used a Poisson regression model to associate the explanatory variables with CRC mortality due to its ability to handle the large number of zeros.

Another limitation was finding accurate sunlight data for each county. Sunlight data that was found was associated with specific airports; however, not every county has an airport. Only Class I airports were used in this analysis because of their data reliability; smaller class airports have less reliable data collection instruments. Class II airport sunlight information was only used for one state (Oklahoma), because Oklahoma contained no Class I airports. Thus, we averaged the sunlight data from two Class II airports to compensate for the lack of the Class I airport. As a result of not having airports in every county, each county within a specific state was assigned the same solar radiation value. Such assignments can result in large discrepancies in states with a large range of latitudes (e.g., California). It could be anticipated that counties in northern California would have less radiation coverage than counties in southern California. Further, a solar plate could not be equally compared to the human body and how fast humans absorb solar radiation.

Another possible limitation is the fact that the education data is calculated for individuals at least 45 years old, instead of at least 50 years old. The incidence of higher education has increased with time, so the population of 45-49 year olds could potentially skew the education data and thus create a limitation within our analysis.

Other factors may have further influenced the results presented in this work. The negative association between living in a rural area and the risk of CRC mortality could occur because individuals are in cleaner air environments, or because of exposure to more sunlight (which has been shown to decrease mortality rates). As shown in Table 2, obesity is highly correlated with diabetes. It is possible that diabetes may be the actual causal factor in colorectal cancer mortality rates. Obesity was chosen as the factor used in the analysis because, through lifestyle changes, individuals would be able to lose weight. This means that there is a way to educate people to decrease the risk of mortality from CRC. Diabetes is a more difficult factor to control. Individuals can regulate and control their insulin levels, but one cannot necessarily cure diabetes (at least not without strong weight regulation).

This data can be used in order to identify risk factors in specific counties and better provide education about how to reduce the risk of and mortality from colorectal cancer. Future studies may want to investigate data that connects specific individuals with their education rates so that it is not rough percentages for the entire county. Also, for the study of rural populations, obesity, and poverty levels, it is valuable to have this data at least for each specific race. Although 45% of a specific population may live in a rural area, this does not mean that this population is equally divided by race. A study that had population data for each race would be beneficial when attempting to conduct separate regression models for each racial group.

**LIST OF TABLES**

Table : List of variable names for the study and their descriptions

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| ETRH | Average extraterrestrial horizontal radiation in watts by square meter by state |
| avg\_Diabetes | Average diabetes rate from 2005-2007 |
| avg\_Obesity | Average obesity rate from 2005-2007 |
| %urban | Percent of total urban population |
| %rural | Percent of total rural population |
| %below\_pLevel | Percent of total population below poverty level |
| med\_income$ | Percent of female population over 50 years of age |
| %pop\_over50 | Percent of total population over 50 years of age |
| %Mpop\_over50 | Percent of male population over 50 years of age |
| %Fpop\_over50 | Percent of female population over 50 years of age |
| %pop45+\_w/HSdegree+ | Percent of total population with at least a high school degree |
| %W\_over50 | Percent Caucasian (non- Hispanic/Latino) population over 50 years of age |
| %H/L\_over50 | Percent Hispanic/Latino population over 50 years of age |
| %AA\_over50 | Percent African-American population over 50 years of age |
| Transform.Rate.AA | Natural log transform of the African-American colorectal cancer rate |
| Transform.Rate. Hispanic | Natural log transform of the Hispanic colorectal cancer rate |
| Transform.Rate.White | Natural log transform of the Caucasian colorectal cancer rate |

Table 2: summary statistics (rounded to three significant digits) for each of the independent and dependent variables used in the regression model. All measures, aside from sample size (n), are measured as rates (%).

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Column** | **n** | **Mean** | **Variance** | **Std. Dev.** | **Std. Err.** | **Median** | **Range** | **Min** | **Max** | **Q1** | **Q3** |
| County.Death.Rate.AA | 3108 | 7.496 | 12323.866 | 111.013 | 1.991 | 0.000 | 3178.218 | 0.000 | 3178.218 | 0.000 | 1.817 |
| County.Death.RateH | 3108 | 1.169 | 103.082 | 10.153 | 0.182 | 0.000 | 333.333 | 0.000 | 333.333 | 0.000 | 0.000 |
| County.Death.Rate.W | 3108 | 1.614 | 1.744 | 1.321 | 0.023 | 1.500 | 46.118 | 0.000 | 46.118 | 1.194 | 1.864 |

Table 3: Pearson’s correlation coefficients for all pairs of variables considered in the study; corresponding p-values provided in parentheses.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Correlation matrix:** | |  |  |  |  |  |  |  |
|  | ETRH | avg.Diabetes | avg.Obesity | %rural | %below.pLevel | med.income$ | %Mpop.over50 | %Fpop.over50 |
| avg.Diabetes | 0.486426 |  |  |  |  |  |  |  |
| (< 0.0001) |
| avg.Obesity | 0.222285 | 0.754106 |  |  |  |  |  |  |
| (< 0.0001) | (< 0.0001) |
| %rural | -0.05908 | 0.265594 | 0.226546 |  |  |  |  |  |
| -0.001 | (< 0.0001) | (< 0.0001) |
| %below.pLevel | 0.371277 | 0.507296 | 0.427024 | 0.183113 |  |  |  |  |
| (< 0.0001) | (< 0.0001) | (< 0.0001) | (<0.0001) |
| med.income$ | -0.19985 | -0.51078 | -0.43799 | -0.44952 | -0.74754 |  |  |  |
| (< 0.0001) | (< 0.0001) | (< 0.0001) | (<0.0001) | (<0.0001) |
| %Mpop.over50 | -0.17663 | 0.120704 | -0.11736 | 0.515155 | -0.09674 | -0.31038 |  |  |
| (< 0.0001) | (< 0.0001) | (< 0.0001) | (<0.0001) | (<0.0001) | (<0.0001) |
| %Fpop.over50 | -0.09379 | 0.286082 | 0.020292 | 0.391136 | -0.03847 | -0.37822 | 0.901017 |  |
| (< 0.0001) | (< 0.0001) | -0.2589 | (< 0.0001) | -0.0323 | (< 0.0001) | (< 0.0001) |
| %pop45+.w/HSdegree+ | -0.43387 | -0.27551 | -0.40144 | 0.160718 | -0.50384 | 0.155454 | 0.754457 | 0.67234975 |
| (< 0.0001) | (< 0.0001) | (< 0.0001) | (< 0.0001) | (< 0.0001) | (< 0.0001) | (< 0.0001) | (< 0.0001) |