International Journal of Mathematical, Engineering and Management Sciences
Vol. 6, No. 3, 944-960, 2021
https://doi.org/10.33889/IJMEMS.2021.6.3.056

Age Dependent Analysis of Colon Cancer Tumours Using Mathematical and Statistical Modelling

Vidya Bhargavi Machavaram
GITAM Institute of Science,
GITAM (deemed to be) University, Visakhapatnam, Andhra Pradesh, India.
Corresponding author: vidya.msc05@gmail.com

Sireesha Veeramachaneni
GITAM Institute of Science,
GITAM (deemed to be) University, Visakhapatnam, Andhra Pradesh, India.
E-mail: vsirisha80@gmail.com

(Received on September 18, 2020; Accepted on March 7, 2021)

Abstract
Colon cancer is the third most commonly diagnosed cancer and the second leading cause of cancer death in men and women combined in the United States. In this work, we performed mathematical and statistical modelling of Tumour sizes as a function of age for four different races. Mathematically, based on the behaviour of the data for each race, we partitioned ages of subjects into several intervals. The mathematical function that characterizes the size of the Tumour as a function of age was determined for each age interval. Statistically, using quantile regression, we designed models that are more robust at specific quantiles using Tumour size and age as dependent and predictor variables.

Keywords- Colon cancer, Line plots, Quantile regression, Statistical modelling, Mathematical modelling.

1. Introduction

Cancer is a sickness characterized through the unchecked division and survival of abnormal cells. When this form of abnormal boom occurs within the colon or rectum, it is known as colorectal cancer (CRC). Combination of colon and rectum is known as large intestine or large bowel. The colon has four sections, the ascending colon starts with the cecum (a pouch wherein undigested food is obtained from the small intestine) and extends upward at the right side of the stomach. The transverse colon is so known as because it travels the body from right to left side. The ascending and transverse colon is collectively referred to as the proximal colon. The descending colon descends at the left aspect. The sigmoid colon, which is known as for its “S” form, is the very last part of the colon and joins the rectum. The descending and sigmoid colon are collectively known as the distal colon (Jacobs et al., 2018).

Colon cancer is the maximum commonly identified cancers and the second main cause of cancer demise in men and women combined inside the United States (Bonsu, 2013). The risk of CRC increases with age. The median age at analysis for colon cancer is 68 in men and 72 in females; for rectal cancer, it is 63 years of age in males and females (Favoriti et al., 2016). As the result of rising CRC incidences in early age groups coincident with declining rates in older age people, the percentage of instances identified in people younger than age 50 accelerated from 6% in 1990 to 11% in 2013 (Few and Edge, 2008). Most of those instances (72%) occur in individuals who are in their 40s. CRC occurrence and mortality costs are highest in non-Hispanic African Americans (NHBs) and lowest in Asians/Pacific Islanders (APIs). During 2009-2013, CRC occurrence rates in African Americans had been about 20% higher than the ones in non-Hispanic Caucasians (NHWs) and 50% better than APIs. The disparity for mortality is two times that for incidence; CRC death quotes in African Americans are 40% higher than in NHWs and double the ones in APIs (Augustus and Ellis, 2018). Reasons for racial/ethnic disparities in CRC are complicated to study but there are noticeable differences in the treatment offered to patients based on their socioeconomic status. According to the US Census Bureau, 24% of African Americans lived in poverty in 2015, in comparison to 11% of Asians and 9% of NHWs (Proctor et al., 2016). In this paper, we modelled Tumour sizes as a function of age for four races by applying both mathematical and statistical techniques. Mathematical/Statistical models have demonstrated to be helpful in establishing definite and clear relationships between systems and stages in cancer (Anderson and Quaranta, 2008; Byrne, 2010). The competency of statistical/mathematical models lies in its capacity to proclaim already obscure or absurd phenomenal principles that are disregarded by a subjective way to deal with science (Altrock et al., 2015). Paterson et al. (2020) developed a stochastic model to estimate the order of procurement of driver mutations that lead to colorectal Tumours, probability of colorectal malignant Tumours. Univariate analysis performed by Pages.et.al (2005) on 959 specimens of resected colorectal cancer proved that the existence or non-existence of histologic signs of early metastatic invasion exhibited significant differences in disease-free and overall survival rate (Pages et al., 2005). DePillis et al. (2013) developed a model to establish relationship between colon cancer growth and treatment using a system of nonlinear ordinary differential equations (ODE).

Bergin, applied quantile regression approach on rural and urban colorectal and breast cancer data. Subjects from rural areas with colorectal cancer had longer intervals at the 50th, 75th and 90th percentiles (Bergin et al., 2018). Researchers, policymakers, and clinicians all use quantile regression. Healthcare expenditures uses this statistical method to interpret observations from various parts of the outcome’s distribution across different quantiles of the distribution (Olsen et al., 2012).

Many research works using mathematical and statistical modelling are published on various cancers. Very few research works are published in the area of age dependent analysis on colon cancer using mathematical and quantile regression. In the first part of this paper, we employ a line graph approach and fit a regression line using ordinary least squares approach to find a model for predicting average Tumour growth as a function of age. This is a preliminary analysis that can shed light for subsequent future research. In the later part, we employ quantile regression approach for modelling Tumour sizes as a function of age. This approach helps to conduct a granular research modelling for the data mainly at the extreme quantiles. Estimates of these regression models can be used to produce forecasts.

2. Data

The data for this research is obtained from SEER database registry (Ratnapradipa et al., 2017). This data source SEER (Surveillance Epidemiology and End Results Program), which is a unique, reliable and essential resource for investigating various cancers.

Colon cancer data from SEER database from 2004 until 2015 is used in this work (Howlader et al., 2016). We pre-processed the information of colon malignant Tumours to expel redundancies and missing data. The resulting data set had 30,251 records. We analyse the rate of change of Tumour growth separately for the four races, Caucasians (88.9%), African Americans (10.4%), American Indians (0.3%) and others (0.4%). 49.7% of the data are male subjects and 50.3% of the data are female subjects with colon cancer. The mean (standard error) of the age of the subjects and Tumour sizes for the entire data are 67.74 (0.08) and 48.14 (0.2) respectively. Detailed data description can be found elsewhere (Bhargavi et al., 2020).

3. Line Plots

The fundamental concept of visually exploring data is to provide the information in some visible shape, permitting the researcher to get insight into the records/data and then infer conclusions. Present descriptive data mining techniques have proven to be of high value in exploratory data evaluation and in addition, they have an excessive capacity for exploring huge databases. Visual representation of data is mainly beneficial when very little is known about the data and when there is no definite suitable technique available to explore the data. Visible representation of the data normally allows a faster exploration and provide results that are more accurate. Especially in instances in which software algorithms fail (Keim, 2002). A line graph is an exploratory data analysis technique, which presents a visual display of the relation between two continuous variables. Line graph is very effective mainly when working with time series data, distance data, etc. They are useful in identifying patterns like trends, and seasonal effects in the data. Connecting values with a line along with time provides true insight of data if (1) the intervals are same in size, (2) the intervals are in right order, and (3) values are recorded at all periods (Koenker and Bassett, 1978). As missing values represent a discontinuity in the information, we processed our initial data to get rid of any missing values. In this work, x-axis of a line graph represents age of the subjects and y-axis represents the average Tumor sizes. For each pair of Age-Tumor size values, a point is marked using the coordinate system. Line graph involves connecting all these points with a line, usually by the method of ordinary least squares (OLS) regression, which may be useful for prediction purposes. The line graph provides a clear display of trend in event such as growth of Tumor sizes along the age axis.

4. Quantile Regression

Quantile regression has better elasticity when analysing data. This method is helpful when there exist differences among the distribution of a dependent variable, such as distinct relationships at different parts of the dependent variable’s distribution (Koenker and Kevin, 2001; Le Cook and Manning, 2013). For instance, quantile regression can be used to determine how many days a treatment should be administered which helps recover 90% of subjects (for 100 subjects with similar conditions), such that there is no interest in the mean.

Quantile regression helps for comprehending relationships between variables outside of the mean of the data, it is useful in understanding outcomes not normally distributed that lack linear relationships with independent variables (Le Cook et al., 2013; Hong et al., 2019). In quantile regression, we estimate the coefficients by minimizing the absolute deviations from the median. Compared to mean, the median is a resistant measure of central tendency that is not affected greatly by outliers.

Recent years have witnessed a constant boom in use of quantile regression in cancer research. A PubMed search returned 103 guides on programs of quantile regression related to most cancers research from 2014 to 2018. Xu et al. (2019) evolved a G-E interaction identification method with the use of the quantile regression method, as most of the existing G-E interplay approaches for analysis records cannot accommodate lengthy-tailed or infected consequences. Koenker and Bassett (1978) formulated quantile regression model for independent data in the year 1978. Yang et al. (2018) developed a new approach for censored quantile regression estimation. There are two main classified groups of quantile regression approaches. They are (a) an inferential method used in quantile regression known as minimization of weighted absolute deviations. Conditional median and a range of other quantile functions are estimated by minimizing asymmetrically weighted absolute residuals and (b) the maximization of a Laplace likelihood.

Let us assume that y_r and $x_{r,}$ (r=1, 2, 3, 4…, n), as the outcome of interest and as predictor variables respectively. The quantile regression model with τ^{th} quantile for the response y_r given x_r is given by

$$Q_{y_i}\,(\tau \mid x_i) = g\,(x_i\,, \beta) \tag{1}$$

Where $Q_{y_i}\,(\cdot) = F{y_i}^{-1}(\cdot)$ is the inverse of cumulative distribution function of y_i given x_i evaluated at τ with 0< τ<1, g (·) is a known function. The coefficient vector β is estimated by minimizing

$$\sum_{i=1}^{n} \rho_\tau\,(y_i - g(x_i, \beta)) \tag{2}$$

where $\rho_\tau\,(\cdot)$ is the check function defined by $\rho_\tau\,(u) = (\tau - I(u<0))u$ and I(·) denotes the indicator function (Huang et al., 2017).

5. Methodology

In this paper, we investigate the socio-demographic affect towards the growth of cancer Tumours. Our response variable is Tumour size. The covariates include age, gender, and race of the cancer patients. Gender variable was not significant. Two-sample t-test showed no statistical significant difference between the Tumour sizes of males and females in the data. Further, we divided the data into four races; Caucasians (R1), African Americans (R2), Asian Indians (R3) and other races (R4). We present relevant statistical models for each of these races.

Line plots for each race were obtained using ordinary least squares (OLS) regression. Considering Age as the independent variable, x, and Tumour size as dependent variable, y, we obtained line plots. Using point process and visual understanding of the behaviour of the scatter plots, we further divided each sample of the data based on age intervals.

The average rate of change of the Tumour size, A(.), over the interval [a, b], is calculated using $\frac{\Delta y}{\Delta x}$. The instantaneous rate, y'(.), is calculated by finding the first order derivative of the best-fit model at the median of the age interval.

Common regression models such as ordinary least squares and logistic regression often assume that the regression coefficients are constant across the population. Methods such as linear regression can estimate the mean of a dependent variable conditional on values from independent variables. Whereas quantile regression is a method of analysis that is used to estimate the conditional median and quantiles of dependent variables. The most common quartiles are Q1 (25th percentile), the median (50th percentile, and Q3 (75th percentile). Moreover, 25% of observations fall below Q1, 50% observations fall below the median, and 75% of observations fall below Q3. Quantile methods allow for researchers to place less focus on common regression slope assumptions, which in turn enables the researcher to describe variations in the distribution of a dependent variable (Le Cook, 2013). Unlike ordinary least squares regression, quantile regression does not make any assumptions on the distribution of dependent variables. In order to get percentile-based rate of growth of the

Tumours, we have modelled equations for each race with the help of quantile regression. Figure 1 depicts the flowchart of this research work.

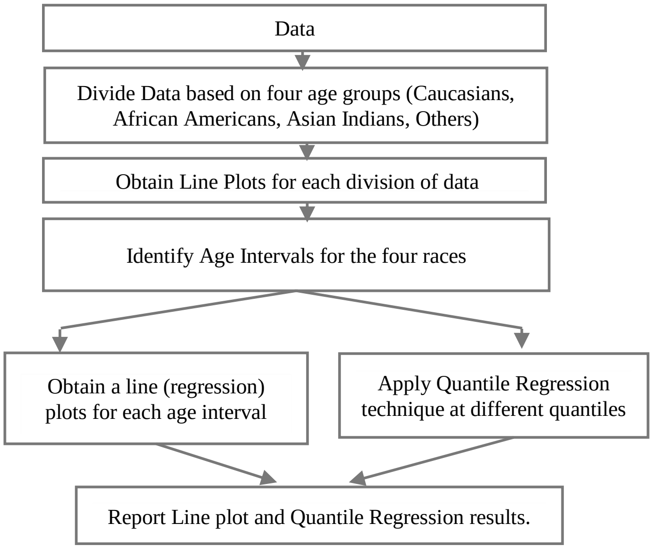
Data
Divide Data based on four age groups (Caucasians, African Americans, Asian Indians, Others)
Obtain Line Plots for each division of data
Identify Age Intervals for the four races
Obtain a line (regression) plots for each age interval
Apply Quantile Regression technique at different quantiles
Report Line plot and Quantile Regression results.

Figure 1. Flowchart of this research work.

6. Line Plots for Each Race in the Data

For each race in consideration, using age as independent variable and average Tumour size as dependent variable, various mathematical equations including logarithmic, linear, quadratic, cubic and exponential were fitted separately. Model with a high coefficient of determination, R2, among the listed ones is chosen as the best-fit model to find out rate of change of Tumour with age. Initially, we plotted a scatter plot of average Tumour size as a function of age for all the races. The graphs for every race is not stationary throughout the age axis. Analysing the data of this form with a single model will not be adequate. We need a specific mathematical model that can handle data of this kind. Thus, we partitioned the age into groups using temporal time point process (Daley and Jones, 2003) while observing changes in the shape of the data and accounting for the quality of the model. Using this process, we arrived at certain age partitions for all the races.

6.1 Line Plot for Average Tumour Size as a Function of Age for Caucasians

In order to fit a best line plot, we generated a scatter plot for average Tumour size as a function of age for all Caucasians. When fitting the line plot for this data, we notice that the trend is not stationary. To address this, we partitioned the age into four age groups. The intervals we obtained are 17-35 years, 36-60 years, 61-90 years, and 91-106 years for fitting best-fit line plots. Figure 2 shows the scatter plot of average Tumour size as a function of age for all Caucasians.

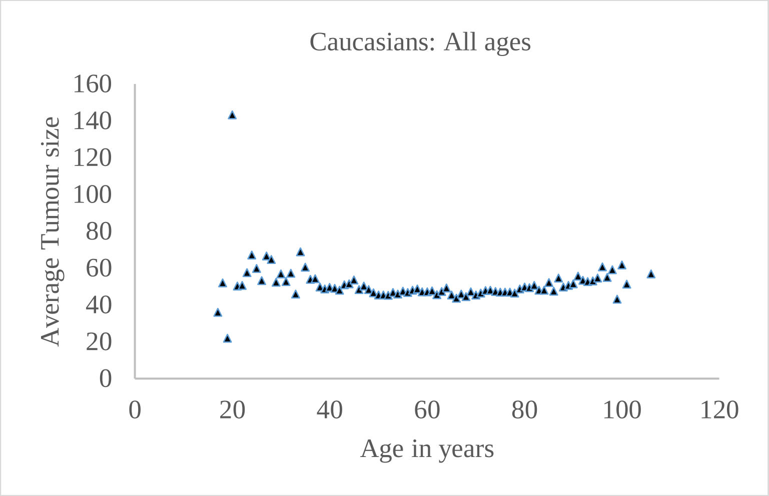

Figure 2. Graph of average Tumour size as a function of age for Caucasians (R1).

Scatter plots for the identified age intervals of Caucasians are in the Figure 3. Once again, for each of these intervals, linear, exponential, quadratic, and cubic models among other models were fitted to the observed data and residual analysis performed. For the age interval of 17-35, a cubic model may seem appropriate for this group with an average residual as -0.0032. Repeating the same process for the rest of the age intervals for all the races and we identified a good-fit model for each such intervals, with a residual mean that is relatively small. Line plot equations for these age intervals of Caucasians are in the Table 1.

Table 1 has the best-fit regression equations along with average rate of change and instantaneous rate of change of Tumour sizes for Caucasians. The average rate of change of Tumour size using, A(.) = [y(b)-y(a)]/(b-a) for each age interval [a, b] and instantaneous rate of change of Tumour size, y'(median) is calculated at the median of the age group. A(.) for the age group of 17-35 is positive and for the rest of the groups it is negative. Tumour size is increasing at an average of 1.3 mm per year in the younger subject group between 17 and 35 years.

Table 1. Regression equations, average rate of change, A(.) and instantaneous rate of change, y'(.), for Caucasians.

Age Intervals	Model	A(.)	y'(.)
All ages	$y = 5E\text{-}05x^3 - 0.0013x^2 - 0.4576x + 69.545$	-	-
[17,35]	$y = 0.049x^3 - 3.9356x^2 + 102.61x - 807.2$	1.2998	-2.6692
[36,60]	$y = 0.0009x^3 - 0.1107x^2 + 4.2458x + 2.094$	-0.0310	-0.1606
[61,90]	$y = -0.0007x^3 + 0.1583x^2 - 12.384x + 362.99$	-0.5984	-0.4512
[91,106]	$y = 0.0087x^3 - 2.5672x^2 + 251.46x - 8145.9$	-0.5603	-1.0497

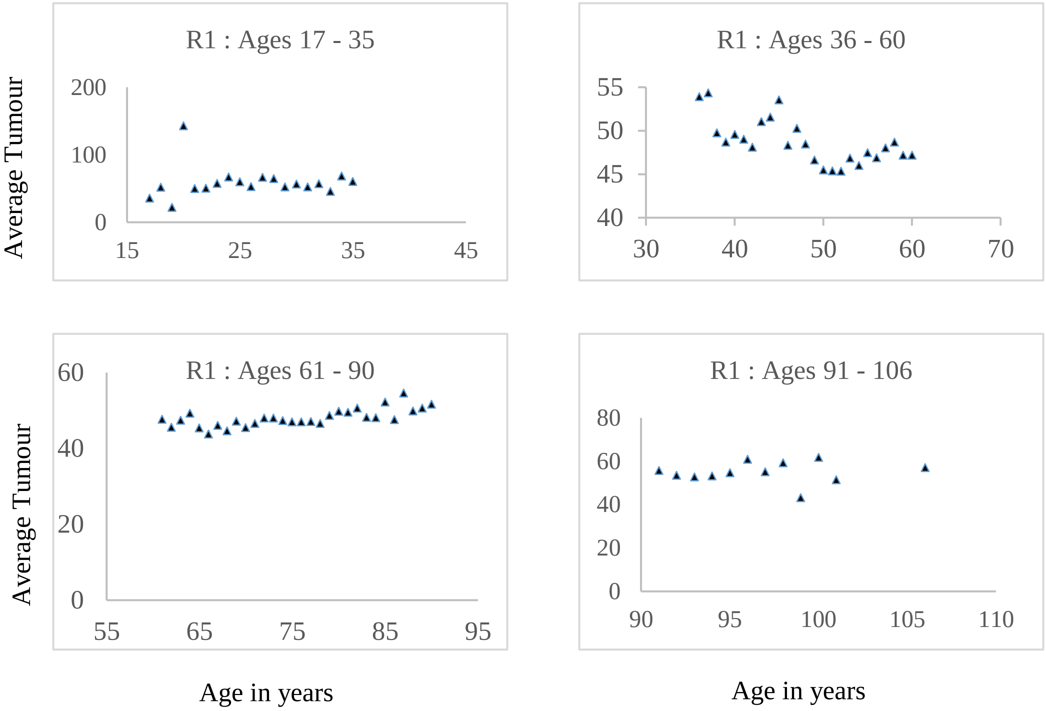

Figure 3. Line plots of age intervals vs. average Tumour sizes of caucasians.

6.2 Line Plot for Average Tumour Size as a Function of Age for African Americans

From the Figure 4, the age intervals we obtained for African Americans are 18-37 years, 38-82 years, and 83-101 years for fitting best-fit line plots.

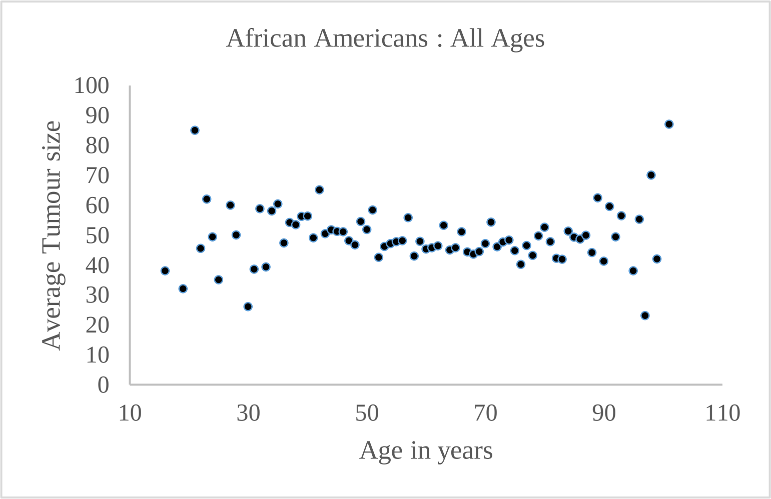

Figure 4. Graph of average Tumour size as a function of age for African Americans (R2).

Table 2. Regression equations, average rate of change, A(.) and instantaneous rate of change, y'(.), for African Americans.

Age Intervals	Model	A(.)	y'(.)
All ages	$y = 0.0002x^3 - 0.0392x^2 + 1.935x + 22.328$	-	-
[18,37]	$y = 0.0302x^3 - 2.4644x^2 + 65.319x - 511.34$	1.0188	-1.7068
[38,82]	$y = -0.0002x^3 + 0.0349x^2 - 2.7164x + 117.15$	-0.7852	3.6316
[83,101]	$y = 0.051x^3 - 13.925x^2 + 1264.7x – 38173$	1.6230	-2.5080

Using the scatter plots for these age intervals as shown in Figure 5, line plots are fitted to understand the trend in these age groups. Table 2 has the best-fit regression equations along with average rate of change and instantaneous rate of change of Tumour sizes for African Americans. A(.) for the age groups of 18-37 and 83-101 is positive and the age group 38-82 value is negative. Tumour size for the younger subject group between 18 and 37 years and older groups of ages 83 and above is increasing at an average of 1.02 mm and 1.62 mm per year respectively.

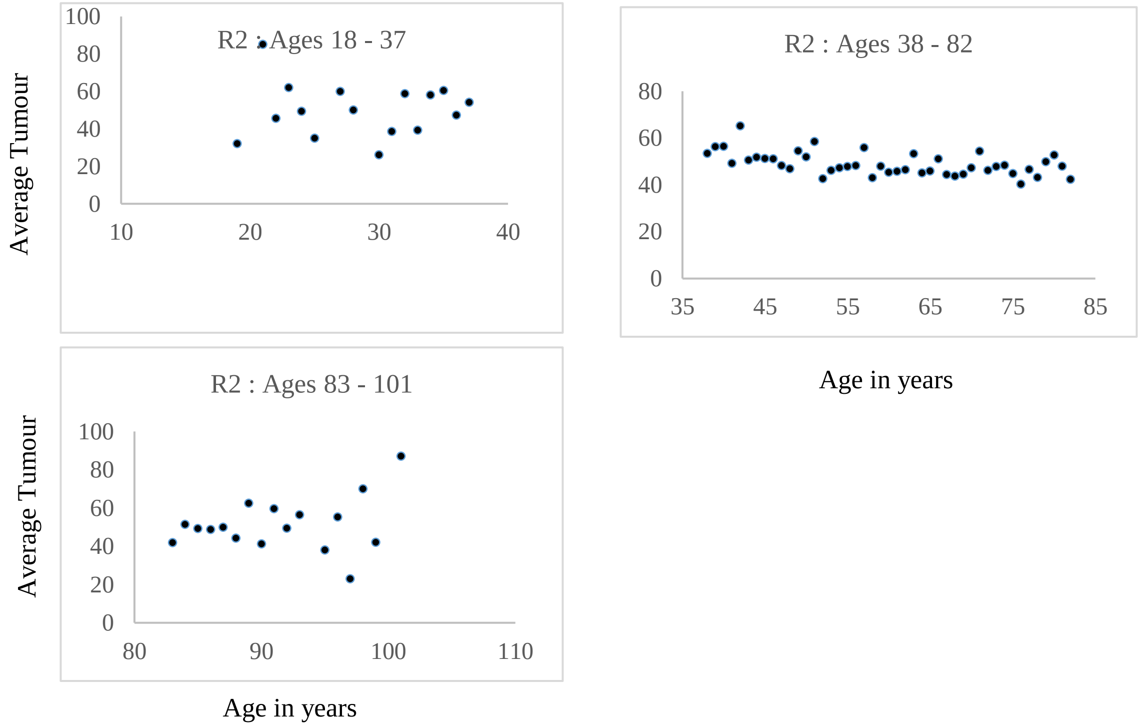

Figure 5. Line plots of age intervals vs. average Tumour sizes of African Americans.

6.3 Line Plot for Average Tumour Size as a Function of Age for Asian Indians

For Asian Indians data, using the scatter plot shown in the Figure 6, we split the age intervals as 27-41 years, 42-62 years, and 63-89 years, respectively.

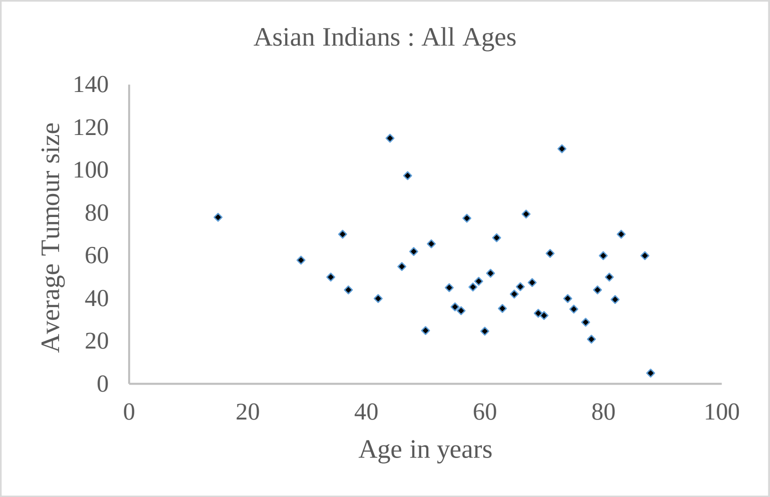

Figure 6. Graph of average as a function Tumour size of age for Asian Indians (R3).

The scatter plots to fit the best-fit line plots for the age groups 27-41 years, 42-62 years, and 63-89 years are in the Figure 7.

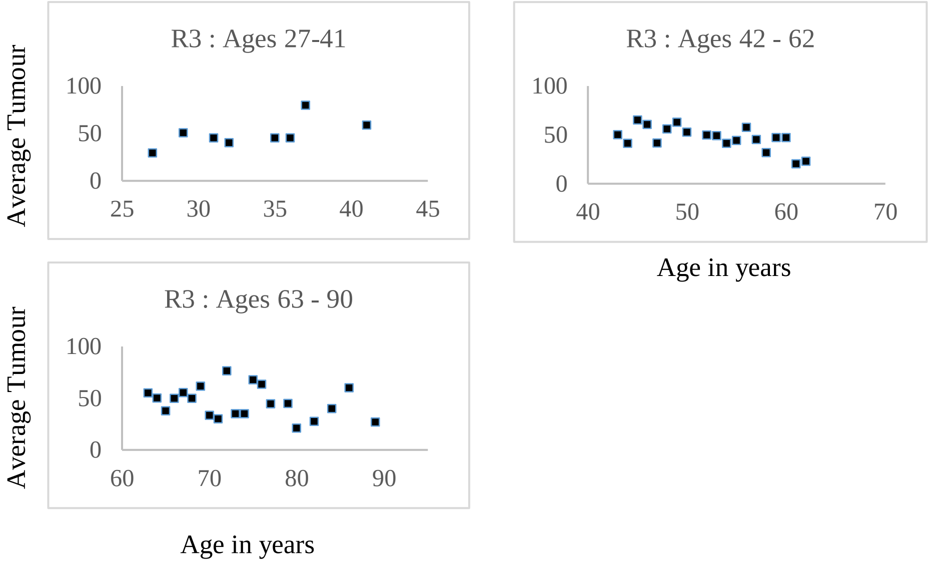

Figure 7. Line plots of age intervals vs. average Tumour sizes of Asian Indians.

After identifying the best-fit line plots for these age intervals, we observe that A(.) for the age groups of 27-41 and 63-89 is positive and the age group 42-62 value is negative. Tumour size for the subject group between 27 and 41 years has an average increase of Tumour size as 2.14 mm per year. Among the four races considered in this study, this is relatively high rate of change of Tumour size.

Table 3 has the best-fit regression equations along with average rate of change and instantaneous rate of change of Tumour sizes for Asian Indians.

Table 3. Regression equations, average rate of change, A(.) and instantaneous rate of change, y'(.), for Asian Indians.

Age Intervals	Model	A(.)	y'(.)
All ages	$y = 0.0004x^3 - 0.0683x^2 + 4.0307x - 23.773$	-	-
[27,41]	$y = 10.93e^{0.0437x}$	2.1438	2.1105
[42,62]	$y = -0.0037x^3 + 0.4185x^2 - 14.763x + 202$	-1.6234	-1.2534
[63,89]	$y = 0.0011x^3 - 0.2638x^2 + 20.952x - 489.33$	0.1011	-0.0848

6.4 Line Plot for Average Tumour Size as a Function of Age for Other Races

Using the scatterplot of other races given in Figure 8, the age intervals are partitioned into 15-42 years, 43-74 years, and 75-87 years for fitting best-fit line plots.

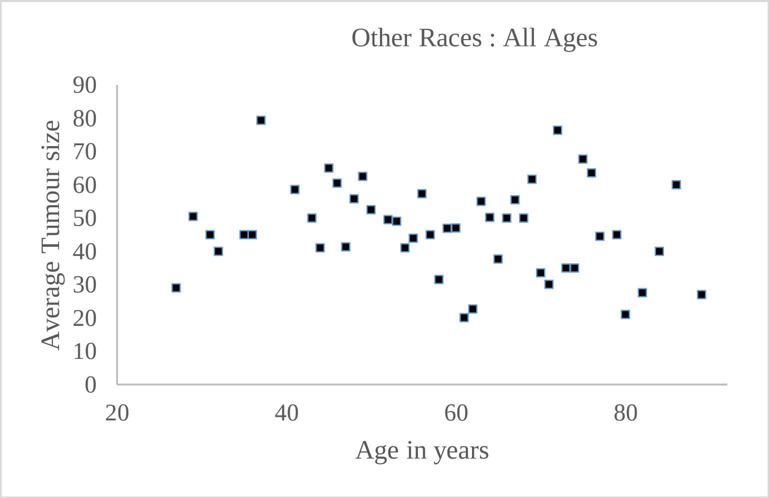

Figure 8. Graph of average Tumour size as a function of age for other races (R4).

Table 4 has the best-fit regression equations along with average rate of change and instantaneous rate of change of Tumour sizes for other races. A(.) for all age groups are negative, indicating that there is a decrease in the average rate of Tumour size per year across all ages for the other races data.

Table 4. Regression equations, average rate of change, A(.) and instantaneous rate of change, y'(.), for other races.

Age Intervals	Model	A(.)	y'(.)
All ages	$y = 90.521e^{-0.011x}$	-	-
[15,42]	$y = -0.0082x^3 + 0.7094x^2 - 20.406x + 252.13$	-1.4460	0.0484
[43,74]	$y = -0.0077x^3 + 1.5384x^2 - 100.15x + 2186.8$	-1.0611	0.7888
[75,87]	$y = -0.2272x^3 + 54.815x^2 - 4400.2x + 117550$	-0.3268	7.8524

The scatter plots used to identify the best-fit line plots for these age groups of other races are in the Figure 9.

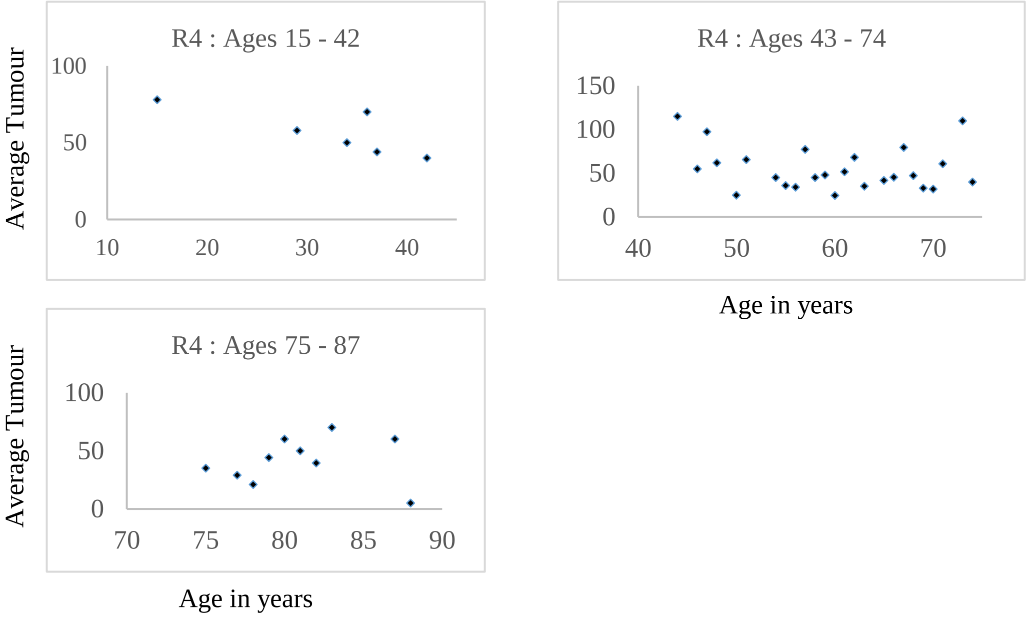

Figure 9. Line plots of age intervals vs. average Tumour sizes of other races.

7. Quantile Regression Approach

Homoscedasticity is one of the important assumptions for ordinary least squares (OLS) regression. In OLS regression, we anticipate that the variance of the residuals must be constant throughout. If the residual variance is not constant and fanned out along, we have a case of heteroscedasticity. In our current data, we find that the distribution of the dependent variable is not normal. Figure 10 shows that as we move left to right, the spread appears to shrink at the very low and high predicted values for Tumour size and with more than a few outliers. We also conclude that there is a pattern in the variance of the residuals violating homoscedasticity assumption of variance for an OLS model. We further performed Breusch-Pagan test for heteroscedasticity. The result of this test reported a test statistic value of 162.208 with a p-value of zero indicating a failure of homoscedasticity. With this support, we proceed to perform quantile regression of the data with Tumour size as the dependent variable, age and race as independent variables.

As linear regression lacks in explaining the effect of a predictor variable like age on the lower and higher Tumour sizes, we consider quantile regression to provide better estimates of age even for lower and higher quantiles of age. We also notice that the data across all age intervals for all the races is either negatively skewed or positively skewed. Quantile regression can help us deal with such behaviour of the data. Quantile regression is more robust to outliers than ordinary least squares regression, and is semiparametric. Quantile regression is also robust for extreme data observations. Using quantile regression, we obtained models for conditional quantiles of Tumour sizes with age, $Q\tau$ (Tumour |age), $\tau \in (0, 1)$.

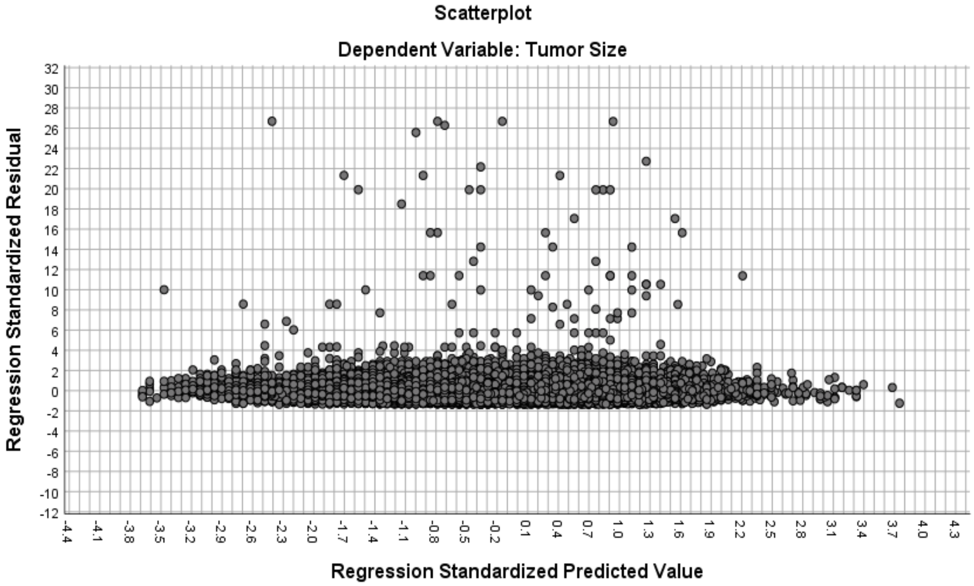

Figure 10. Two-way scatter plot of standardized residuals from the OLS regression.

Quantile regression models for the quantiles ranging from 0.05 until 0.97 are developed. The age parameter did not show any significant effect for the quantiles above 0.3 and below 0.9. This means that the conditional quantile at any level in that range cannot be estimated precisely for the Tumour size prediction as a function of age. Therefore, we focused on the lower quantile levels of 0.3 and below and higher quantiles above 0.9. Note that 0.3 quantile and below represents the age for early cancerous subjects while the 0.9 quantile and above corresponds to the older age cancerous subjects.

Figure 11 documents the behaviour of the parameter estimates using quantile regression. For the independent variables, note that the extreme quantiles are the ones where the nonlinear effect is prominent. The left column plots of the Figure 11 are the intercept term of the model. It shows the predicted Tumour size for each quantile, if there is no increment in the age of the subject. The curve is monotone upwards.

However, our main interest is in the graphs is the middle column, age. Age of the subjects clearly makes much more difference in the lower and upper quantiles. At the 0.05 quantile, a unit increase of the age relates to about 0.1 mm increase in the Tumour size. Similarly, at the upper quantiles, say at 0.92 quantile, Tumour size increases by 0.05 mm. It is interesting to notice that 0.92 quantile of age parameter is significant followed by 0.95 being insignificant and 0.97 quantile as significant.

The right most column of the Figure 11 are the parameter coefficients of the race term. We notice that race is not behaving different from OLS prediction.

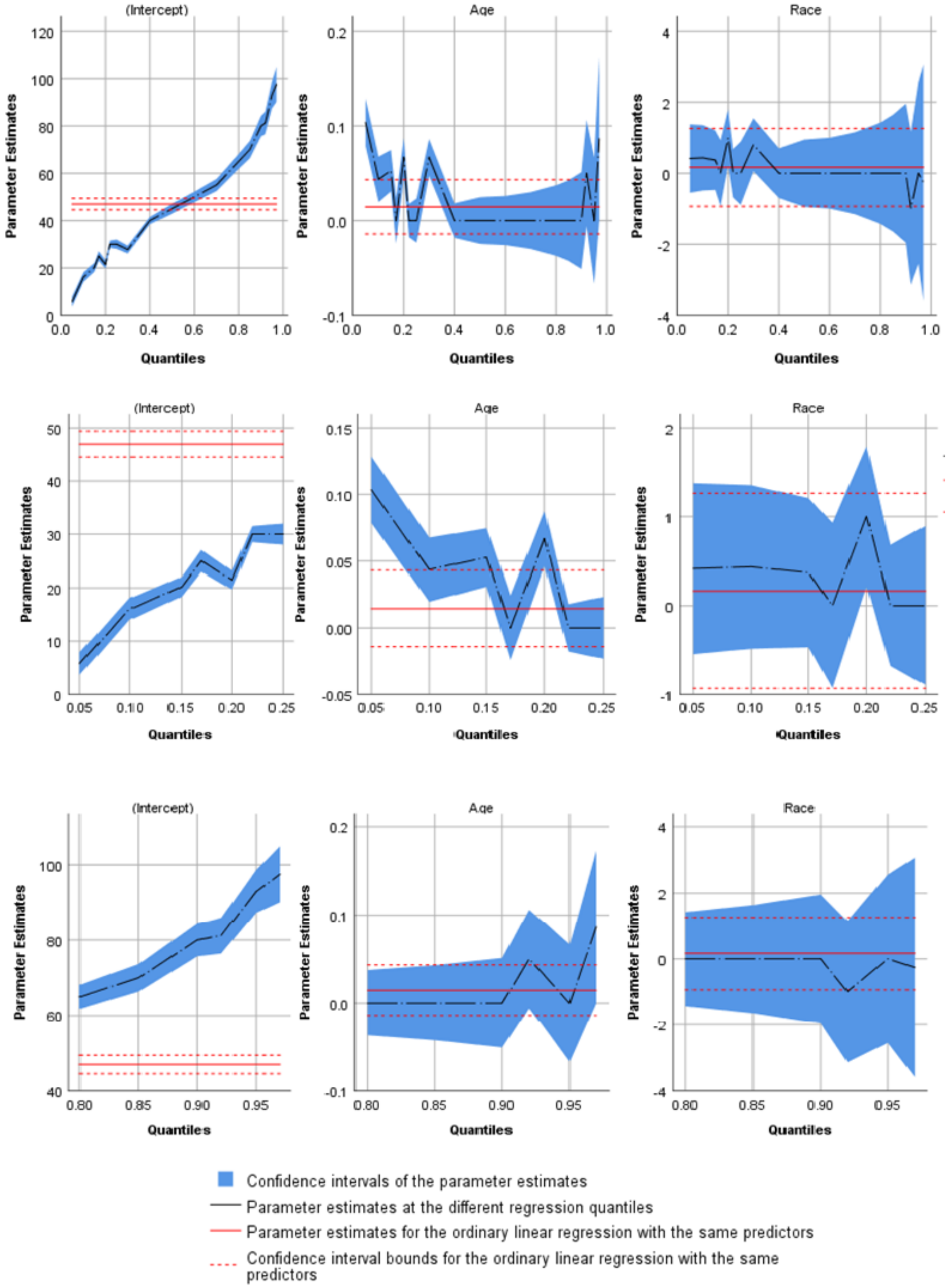

Figure 11. Overall, lower quantiles and higher quantile parameter estimates of age and race.

Compared to all quantiles, it is interesting to observe that age and race are both significant only for 0.2 and 0.3 quantiles. This reveals that age and race factors played an important role in both quantiles. In addition, the effect of age is highly significant and with increase of one year in age is associates with 0.07 mm increase of Tumour size. At the higher quantiles (0.92, 0.95, 0.97), even though the race values were not statistically significant, the opposite signs of the coefficients at these quantiles indicate that the effects of the race variable are heterogeneous. The effect of age was not significant at the quantile 0.95, indicating that age may have heterogeneous effects on Tumour size. Table 5 reflects the modelling results of quantile regression for the lower and higher quantiles. We omit the quantile regression output for the intermediate quantiles, as it is not worth discussing.

Table 5. Quantile parameter estimates (Est), standard error (SE), 95% confidence interval for the covariates (95% CI). Estimates in bold indicate statistically significant effects.

τ		Intercept	Age	Race
0.05	Est	**5.793**	**0.103**	0.414
	SE	1.0869	0.0128	0.4906
	95% CI	(3.663,7.923)	(0.078,0.129)	(-0.548,1.375)
0.1	Est	**16.126**	**0.044**	0.433
	SE	1.0396	0.0122	0.4692
	95% CI	(14.088,18.163)	(0.02,0.068)	(-0.486,1.353)
0.15	Est	**20.158**	**0.053**	0.368
	SE	0.9477	0.0111	0.4277
	95% CI	(18.3,22.015)	(0.031,0.074)	(-0.47,1.207)
0.2	Est	**21.467**	**0.067**	**1**
	SE	0.8942	0.0105	0.4036
	95% CI	(19.714,23.219)	(0.046,0.087)	(0.209,1.791)
0.3	Est	**27.733**	**0.067**	**0.8**
	SE	0.8428	0.0099	0.3804
	95% CI	(26.081,29.385)	(0.047,0.086)	(0.054,1.546)
0.92	Est	**81.183**	**0.05**	-0.983
	SE	2.4344	0.0286	1.0988
	95% CI	(76.412,85.955)	(-0.006,0.106)	(-3.137,1.17)
0.95	Est	93	-3.30E-16	2.76E-15
	SE	2.8918	0.034	1.3052
	95% CI	(87.332,98.668)	(-0.067,0.067)	(-2.558,2.558)
0.97	Est	**97.5**	**0.086**	-0.259
	SE	3.7606	0.0442	1.6974
	95% CI	(90.129,104.871)	(-0.001,0.173)	(-3.586,3.068)

In quantile regression, predicted Tumour size (Q) is the dependent variable. The independent variables are age, and race. The model equations using quantile regression for 0.1 and 0.92 quantiles are in the Equations 3-4:

$$\mathrm{Q}(\tau = 0.1 |\ \mathrm{Age}, \mathrm{Race}) = 16.126 + 0.044\ (\mathrm{Age}) + 0.433(\mathrm{Race}) \tag{3}$$

$$\mathrm{Q}(\tau = 0.92 |\ \mathrm{Age}, \mathrm{Race}) = 81.183 + 0.05\ (\mathrm{Age}) - 0.983(\mathrm{Race}) \tag{4}$$

8. Conclusions

In this paper, we performed age dependent analysis of colon cancer Tumours using mathematical and statistical modelling. We modelled line plots using ordinary least square regression and quantile regression equations for Tumour size as dependent variable with age and race as independent variables. Linear regression equations are developed to each of the four races, Caucasians, African Americans, Asian Indians and other races. We obtained different age intervals for each race and fitted ordinary least squares regression equations. Average rate of change and instantaneous rate of change of Tumour size for each age interval are calculated.

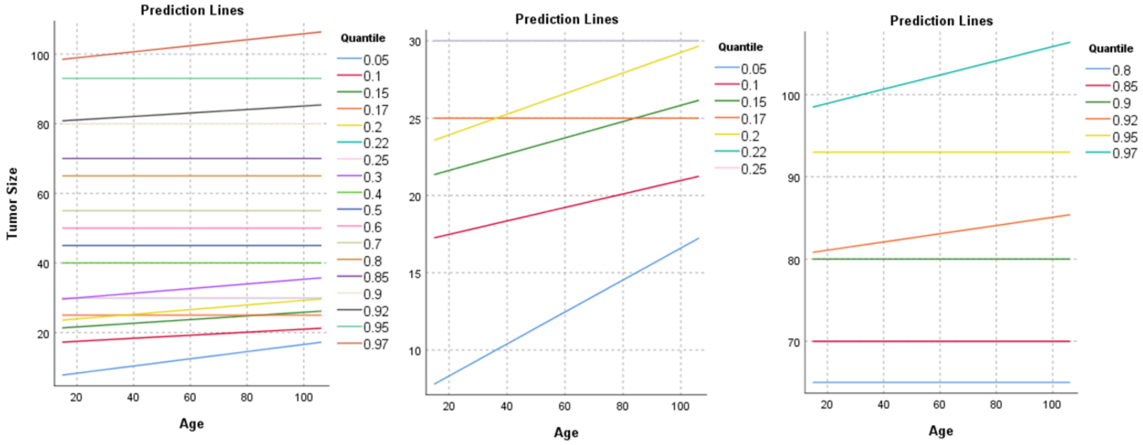

Figure 12. Quantile regression prediction lines of Tumour size.

Caucasians has highest and lowest average rate of change of Tumour size at the age intervals [17, 35] and [61, 90] with average rate values of 1.2998 and -0.5984 respectively. African Americans has highest average rate of change of Tumour size at the age interval of [83, 101] with a value of 1.623. Among all the races considered, Asian Indians has highest average rate of change of Tumour size at the age interval of [27, 41] with a value of 2.1438. Other races are the only race that reported a decline of size of Tumours for all the age groups. However, the instantaneous rate of change for the age group [75, 87] reported a highest value of 7.85 compared to all other races.

Further, in the work, we fitted quantile regression equations with age and race as independent variables and Tumour size as the dependent variable. Quantile regression is a semiparametric tool capable of modelling heterogeneity in the relationship between a dependent and one or more independent variables without making parametric assumptions on the conditional distributions. In this article, we modelled estimating Tumour size of colon cancer using age and race as covariates. The quantile parameter estimates are nonlinear at the extreme quantiles. We modelled quantile regression equations for the extreme quantiles at 0.05, 0.1, 0.15, 0.2, 0.3, 0.92, 0.95 and 0.97. Race variable was significant only for lower quantiles 0.2 and 0.3. Age variable is significant at both the extremes. The estimates for the regression coefficients along with standard errors as well as the 95% confidence interval bounds are obtained. The prediction lines for quantile regression are generated and are presented in Figure 12. Even though this is just a preliminary study, we find clear indications that age plays a very important role in the growth of identified Tumours, mainly when the subjects are identified with colon cancer in their young age.

Conflict of Interest

The authors confirm that there is no potential conflict of interest to publish the paper in the journal.

Acknowledgements

We would like to thank Stanley College of Engineering-Hyderabad, GITAM (deemed to be) University-Visakhapatnam, our families and colleagues for their unconditional support in completing this research article.

References

Altrock, P.M., Liu, L.L., & Michor, F. (2015). The mathematics of cancer: integrating quantitative models. *Nature Reviews Cancer*, *15*(12), 730-745.

Anderson, A.R., & Quaranta, V. (2008). Integrative mathematical oncology. *Nature Reviews Cancer*, *8*(3), 227-234.

Augustus, G.J., & Ellis, N.A. (2018). Colorectal cancer disparity in African Americans: risk factors and carcinogenic mechanisms. *The American Journal of Pathology*, *188*(2), 291-303.

Bergin, R.J., Emery, J., Bollard, R.C., Falborg, A.Z., Jensen, H., Weller, D., Menon, U., Vedsted, P., Thomas, R.J., Whitfield, K., & White, V. (2018). Rural-urban disparities in time to diagnosis and treatment for colorectal and breast cancer. *Cancer Epidemiology and Prevention Biomarkers*, *27*(9), 1036-1046.

Bhargavi, M.V., Mudunuru, V.R., & Veeramachaneni, S. (2020). Colon cancer stage classification using decision trees. In: Raju, K.S., Senkerik, R., Lanka, S.P., Rajagopal, V. (eds.) *Data Engineering and Communication Technology*. Springer, Singapore, pp. 599-609.

Bonsu, N.O. (2013). Age dependent analysis and modelling of prostate cancer data. Tampa: University of South Florida Scholar Commons.

Byrne, H.M. (2010). Dissecting cancer through mathematics: from the cell to the animal model. *Nature Reviews Cancer*, *10*(3), 221-230.

Daley, D.J., & Jones, D.V. (2003). *An introduction to the theory of point processes: elementary theory of point processes*. Springer.

DePillis, L.G., Savage, H., & Radunskaya, A.E. (2013). Mathematical model of colorectal cancer with monoclonal antibody treatments. arXiv preprint arXiv:1312.3023.

Favoriti, P., Carbone, G., Greco, M., Pirozzi, F., Pirozzi, R.E.M. & Corcione, F. (2016) Worldwide burden of colorectal cancer: a review. *Updates in Surgery*, *68*(1), 7-11.

Few, S., & Edge, P. (2008). Line graphs and irregular intervals: an incompatible partnership. *Visual Business Intelligence Newsletter*, *12*(11), 16-29.

Hong, H.G., Christiani, D.C., & Li, Y. (2019). Quantile regression for survival data in modern cancer research: expanding statistical tools for precision medicine. *Precision Clinical Medicine*, *2*(2), 90-99.

Howlader, N., Noone, A.M., Krapcho, M. (2016). SEER cancer statistics review. Bethesda, *MD: National Cancer Institute*, 1975-2013.

Huang, Q., Zhang, H., Chen, J., & He, M. (2017). Quantile regression models and their applications: a review. *Journal of Biometrics & Biostatistics*, *8*(3), 2155-6180.

Jacobs, D., Zhu, R., Luo, J., Grisotti, G., Heller, D.R., Kurbatov, V., Johnson, C.H., Zhang, Y., & Khan, S.A. (2018). Defining early-onset colon and rectal cancers. *Frontiers in Oncology*, *8*, 504.

Keim, D.A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, *8*(1), 1-8.

Koenker, R., & Bassett, Jr.G. (1978). Regression quantiles. *Journal of the Econometric Society*, *46*(1), 33-50.

Koenker, R., & Kevin, F.H. (2001). Quantile regression. *Journal of Economic Perspectives*, *15*(4) 143-156.

Le Cook, B., & Manning, W.G. (2013). Thinking beyond the mean: a practical guide for using quantile regression methods for health services research. *Shanghai Archives of Psychiatry*, *25*(1), 55-59. doi:10.3969/j.issn.1002-0829.2013.01.011.

Olsen, C.S., Clark, A.E., Thomas, A.M., & Cook, L.J. (2012). Comparing least-squares and quantile regression approaches to analyzing median hospital charges. *Academic Emergency Medicine*, *19*(7), 866-875.

Pages, F., Berger, A., Camus, M., Sanchez-Cabo, F., Costes, A., Molidor, R., Mlecnik, B., Kirilovsky, A., Nilsson, M., Damotte, D., Meatchi, T., Bruneval, P., Cugnenc, P., Trajanoski, Z., Fridman, W., Galon, J. (2005). Effector memory T cells, early metastasis, and survival in colorectal cancer. *New England Journal of Medicine, 353*(25), 2654-2666.

Paterson, C., Clevers, H., & Bozic, I. (2020). Mathematical model of colorectal cancer initiation. In *Proceedings of the National Academy of Sciences. Cold Spring Harbor Laboratory, bioRxiv*.

Proctor, B.D., Semega, J.L., Kollar, M.A. (2016). Income and poverty in the United States: 2016. U.S. Government Printing Office, Washington. DC: U.S. Census Bureau.

Ratnapradipa, K.L., Lian, M., Jeffe, D.B., Davidson, N.O., Eberth, J.M., Pruitt, S.L., & Schootman, M. (2017). Patient, hospital, and geographic disparities in laparoscopic surgery use among surveillance, epidemiology, and end results–medicare patients with colon cancer. *Diseases of the Colon & Rectum, 60*(9), 905-913.

Yang, X., Narisetty, N.N., & He, X. (2018). A new approach to censored quantile regression estimation. *Journal of Computational and Graphical Statistics, 27*(2), 417-425. doi: 10.1080/10618600.2017.1385469.

Xu, Y., Wu, M., Zhang, Q., & Ma, S. (2019). Robust identification of gene-environment interactions for prognosis using a quantile partial correlation approach. *Genomics, 111*(5), 1115-1123.