Using Survival Analysis to Build Colon Cancer Survival Time Models

Group 6

March 16, 2023

Gawing Vong [pmygv1@nottingham.ac.uk]
Joseph Cooke [pmyjc23@nottingham.ac.uk]
Oliver Kerr [pmyok4@nottingham.ac.uk]
Xinyu Ma [biyxm5@nottingham.ac.uk]
Xiangyu Liu [smyxl6@nottingham.ac.uk]
Zihang Yu [smyzy4@nottingham.ac.uk]

Abstract

This report investigated the most important prognostic variables associated with the survivability of colon cancer patients and has subsequently devised statistical models for survival time. The methods employed in this report include non-parametric methods such as the Kaplan-Meier estimate and the Log-rank test; semi-parametric methods such as the Cox Proportional Hazards Regression model; and parametric methods such as the Exponential model and the Weibull model. We concluded that the anatomical subsite and the clinical stage of cancer patients at the time of diagnosis are the two most important variables in terms of prognosis for survival using different methodologies. Finally, this allowed us to make survival time predictions using the models we devised for an individual diagnosed with colon cancer.

2023.3

Contents

1 Introduction

Colon and rectal cancers are currently the third most frequent type of cancer globally. As part of the large intestine, the colon is a lengthy tube-like organ that is responsible for drawing out nutrients and water from digested food. Along with lung, prostate, and breast cancer, colon cancer is one of the most common tumours worldwide and is ranked among the major killers. Each year, around 250,000 new colon cases are identified in Europe, making up about 9% of all cancers *(Labianca, Beretta et al. 2010)*.

Modelling the survival times of patients is an essential tool in examining the impacts and characteristics of colon cancer. Predicting the survival times of patients based on selected prognostic variables can assist in the treatment of cancer, by informing healthcare professionals about high-risk individuals who could benefit from more rigorous surveillance and treatment. More specifically, by identifying the factors that are associated with shorter survival periods, more intense treatments can be provided to those high-risk patients, optimising their chances of recovery. By producing efficient survival time models, insights can be made about how and when treatments should be implemented and how long patients should expect to live based on their prognostic factors. Previous studies have been working on creating categorization schemes for staging colon cancer. These systems are intended to help doctors stratify patients based on projected predicted survival time in order to help them choose the best treatments, calculate prognosis, and assess cancer prevention strategies *(O'Connell, Maggard et al. 2004)*.

This report aims to investigate the prognostic variables associated with colon cancer using methods from survival analysis. Statistical models will be produced that model the dependent variable (survival time) using selected prognosis variables as independent variables. More specifically, the report will begin by introducing the variables in the dataset in section 1.1, giving a brief description of their meaning and how they have been interpreted for the analysis. In section 2, we looked at possible methods for survival analysis which include the Kaplan-Meier estimate, log-rank test, Cox regression model, exponential model and finally the Weibull model. In section 3, we found the results of these methods which define anatomical subsite and clinical stage at diagnosis of a patient as the most important variables related to survival time. The Cox regression model and the Weibull model produce predictions of the survival time of a patient which we assess the fit for using statistical techniques such as maximum likelihood estimation. Section 4 summarised the main findings from each section and highlighted key models which predict survival time for colon cancer patients. Section 5 detailed the clear limitations involved with our study helping to guide future research into this topic. Section 6 gave references which have been used in this report. Finally, section 7 accentuates any extra plots which might be of interest.

1.1 Description of Variables

Data has been collected from 15,564 colon cancer patients in a northern European country from 1975 to 1994. The data consists of variables which have been described in Table 1 below:

Table 1: Description of variables

Explanatory Variable	**Description**
Survival Time	The survival time of colon cancer patients is given in both months (surv_mm) and years (surv_yy). This records the length of time a patient is alive for after their colon cancer diagnosis. The survival time of patients is the dependent variable of the study.
Sex	The sex of the patient. Male or Female.
Age	The age of the patient at diagnosis in years.
Clinical Stage (stage)	The stage of cancer at the time of diagnosis. This is an indicator of how far the cancer has progressed and is a factor variable. The status is 'unknown', 'localised', 'regional' or 'distant'. Localised refers to when the cancer has not spread within the body and is still confined to its area of origin. If the cancer is 'regional' it has spread from its origin to another part of the body; with colon cancer, common secondary locations are the liver and the lymph nodes. A 'distant' diagnosis refers to when the cancer has spread to a part of the body that is not directly adjacent to its area of origin.
Time of Diagnosis	The month (mmdx) and the year (yydx) of the diagnosis.
Vital Status (status)	The vital status of the patient at the last date of contact is also considered. Patients are categorised as 'alive', 'died of cancer', 'died with cancer' or 'lost to follow-up'. 'Died of cancer' refers to a patient who died from cancer, whereas 'Died with cancer' refers to any patient who died whilst having cancer but from an alternative cause. Patients 'lost to follow-up' might be due to unwillingness to continue in the study or difficulties making contact to retrieve the necessary information.
Anatomical Subsite of Tumour (subsite)	The subsite of a cancer tumour is where the tumour is located. The colon cancer data for this factor variable is split into four different subsite types: "Transverse", "Coecum and ascending", "Descending and sigmoid" or "Other or Not Otherwise Specified".
Respiratory Status (resp)	The factor variable considering whether patients have respiratory complications at entry.

2 Methods

This section will discuss definitions, assumptions and statistical methods which have been used in the survival analysis of colon cancer data. Hence this allows us to discuss the significance of each of the previously outlined variables on survival time. As described above there were also patients who are categorised as 'lost to follow up'. It should be noted that our data set included 4 of these observations and they were not considered in the analysis as they would not have significant effects on the final results.

2.1 Definitions in Survival Analysis

2.1.1 Survivor functions and Hazard functions

The survivor function, denoted by $S(t)$, is a probability function that represents the probability of an individual surviving beyond a time t. The function is formulated as follows:

$$S(t) = P(T > t) = 1 - F(t)$$

where T is a random variable representing the time until an event occurs, and $F(t)$ is the cumulative distribution function (CDF) of T.

The hazard function, denoted by h(t), is a measure of the instantaneous rate of failure at time t, given that the individual has survived up to a time t. The function is defined as:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \times \frac{1}{S(t)} = \frac{f(t)}{S(t)}$$

where f(t) is the probability density function (PDF) of T. Note that h(t) is alternatively known as a conditional failure rate and also as a hazard rate.

The cumulative hazard function, denoted H(t), is the collected risk of experiencing the event of death at a time t. This function is defined as:

$$H(t) = \int_0^t h(x)dx$$

and the survivor function is defined using the cumulative hazard function as:

$$S(t) = exp(-\int_0^t h(x)dx) = exp(-H(t))$$

where $H(t)$ is the cumulative hazard function.

2.1.2 Censored Data

Censored data is any data for which the exact event time is unknown. Our statistical models and methods for prediction must account for a proportion of the data set which is right-censored. A 'right-censored' patient is someone who was either alive at the last contact time before the study is concluded, or prematurely withdrew from the study. In our data set, those who are in the alive category were assigned a survival time corresponding to the amount of time from diagnosis until the date that the study concluded; therefore, we do not know their true survival time, and must consider them as censored. Those patients who 'Died of Cancer' or 'Died with Cancer' are all considered as not censored. Out of 15,564 patients in this study, 4,642 were recorded as being alive at the last point of contact, accounting for approximately 29.8% of the data set. All of the methods used in our analysis are techniques specifically designed for survival analysis and automatically take into account the censoring problem.

2.2 Non-parametric Methods

2.2.1 The Kaplan-Meier Estimate

The Kaplan-Meier estimate is implemented to measure the proportion of patients living for a given amount of time after a cancer diagnosis. Using this estimate, Kaplan-Meier survival curves can be obtained for groups of patients with differing prognostic factors to examine any statistical differences in their survival times. The estimate involves calculating the probability of the event of interest occurring at a given point in time. The survival probability at any given period of time is given by the formula:

$$\widehat{S}(t) = \prod_{i:t_i \leq t} (1 - \frac{d_i}{n_i})$$

where $\widehat{S}(t)$ denotes the Kaplan-Meier estimate of survivor function; t_i denotes the time at which at least one death occurred; d_i denotes the number of deaths that occurred at time t_i; and n_i denotes the number of individuals known to have survived up to time t_i *(Collett 2015).*

The total probability of survival within the given time period under consideration is the product of all the survival probabilities at the time periods that precede this time. To give an example, the probability of a patient surviving 6 months is the probability of the patient surviving the sixth month given that they survived the previous 5 months.

Kaplan-Meier curves use the Kaplan-Meier estimator to estimate the survivor function of patients. The horizontal axis denotes the length of time since the cancer diagnosis (in months) and the vertical axis is a value between 0 and 100 which describes the percentage of patients expected to survive a given length of time. Curves commonly include a shaded region which denotes the 95% confidence interval. Kaplan-Meier curves can be used to identify differences between two groups of subjects, for example, those patients with different colon cancer subsites. There are two main differences that can be explored in curves, a vertical gap between two curves means that one group had higher expected survival prospects at some given point in time. A horizontal gap has a slightly different interpretation where for one of the groups, it has taken longer to experience a certain proportion of deaths.

Figure 1 shows a Kaplan-Meier curve describing the estimated survival rates of all patients in the colon cancer study. The plot shows how the survival rate of all patients seems to fall sharply in the initial months after the diagnosis and then continues to fall throughout the period but at a slower rate. After 1.5 months, 88.4% of patients survive but just after a year (12.5 months) this falls to 63.2% of patients. The survival rate of patients just after 6 years (73.5 months) is 33.6%. There is not much variation in this plot of all patients throughout the whole time period, however, it increases slightly for longer periods since diagnosis especially from 200 months onwards. According to the plot, the expected median survival time for all colon cancer patients is 27.5 months.

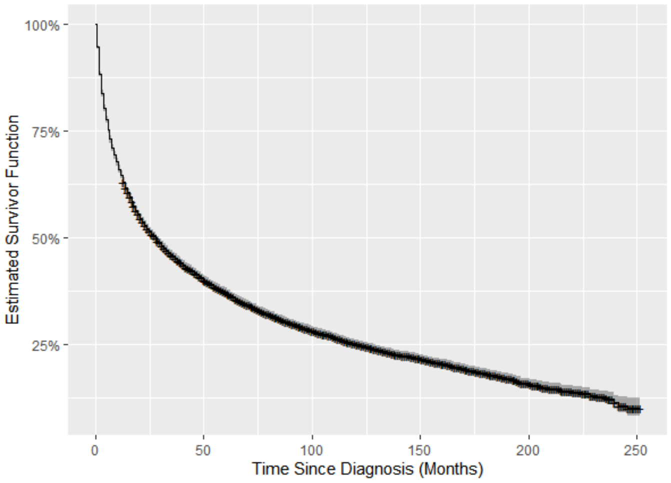


Figure 1: Kaplan-Meier Curve for Training Set

2.2.2 Log-rank test

The log-rank test is a statistical method used to compare the distribution of survival time in two or more groups. This method can be used to test if there are differences in survival distributions among different levels of factor variables, providing statistical confirmation of the differences in Kaplan-Meier curves. The null hypothesis tested by the log-rank test assumes the survival distributions of different groups are equal. The rejection of the null hypothesis means that the survival rates statistically differ among varying groups. Two formulas can be used to calculate log-rank statistics which are defined below:

$$\chi_1^2 = \frac{(|O_1 - E_1| - 1/2)^2}{E_1} + \frac{(|O_2 - E_2| - 1/2)^2}{E_2}$$

$$\chi_1^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

where E_1 and E_2 represent the expected number of deaths in the first and second groups respectively; O_1 and O_2 represent the actual number of deaths in the first and second groups respectively. Similarities are found in the two formulas since they both test whether the survival distribution in two groups is identical and whether the calculated log-rank statistic approximates a χ^2 distribution with 1 degree of freedom. However, the first formula is a formula for log-rank statistics with continuity correction. This means it makes the test statistic more conservative. When sample sizes are moderate or large, we can omit the use of continuity correction. The second formula is the most common form of the log-rank statistic.

When the number of groups is more than two, we need this new formula to calculate the test statistic:

$$\chi_{g-1}^2 = \sum \frac{(O_g - E_g)^2}{E_g}$$

In this new formula, O_g and E_g represent the observed and expected death in group g respectively. The test statistic approximates a χ^2 distribution with a $g - 1$ degree of freedom.

2.3 Semi-parametric Method: Cox regression model

Both non-parametric methods above, the Kaplan-Meier curves and log-rank tests, are limited to analyzing one predictor or comparing multiple categories of survival time, and are not suitable for considering the impact of factors together. To address these limitations and simultaneously assess several predictors, including quantitative and categorical ones, the proportional hazards regression model, also known as the Cox regression model, can be used. One objective in modelling survival data is to determine which combination of variables affects the form of the hazard function. Another reason to model the hazard function is to obtain an estimate of the hazard function for an individual. This allows us to form a relationship between the survivor function and the hazard function by estimating the baseline hazard function which is our objective in using this method. We can use this to find the median and mean survival time which can be estimated for future patients. Formulas and interpretations will be given below.

If the hazard rates of n individuals at a particular time depends on the values $\mathbf{x} = (x_1, x_2, ..., x_p)$ for p explanatory variables $\mathbf{X} = (X_1, X_2, ..., X_p)$, then the Cox model defines the hazard of death at time t for the ith individual:

$$h_i(t|\boldsymbol{\beta}, \mathbf{X_i}) = h_0(t)exp(\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi})$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of regression parameters unrelated to time, and $h_0(t)$ is called the baseline hazard which represents the hazard of an individual with $\mathbf{x} = \mathbf{0}$.

In this circumstance, the hazard functions of two individuals with covariates $\mathbf{X_a}$, $\mathbf{X_b}$ are at constant proportion at all times, because

$$\frac{h(t|\mathbf{X_a})}{h(t|\mathbf{X_b})} = exp(\beta(\mathbf{x_a} - \mathbf{x_b}))$$

which gives the proportional hazards. Thus, we can ignore $h_0(t)$ to estimate β, if we are not primarily concerned with the precise form of the hazard, but with the effects of the covariates.

However, some assumptions should be satisfied in the Cox regression model:

1. Proportional hazards (PH) assumption: For different levels of the predictor variables, the hazards (i.e. the risk of an event occurring at a given time) are proportionate throughout time. In other words, the risk ratio between any two groups remains constant over time.

2. Independence assumption: The observations are independent hence unrelated to one another. This indicates that one person's survival time is independent of another person's survival time.

3. Non-censoring assumption: Censoring happens at random and is unrelated to the relevant event. In other words, there is no correlation between the probability of censoring and the probability of the event happening.

4. Linearity assumption: The log-hazard ratio is linearly related to predictor variables. Using diagnostic charts such as the martingale residuals plot, this assumption can be verified.

5. Adequate sample size: The sample size should be large enough to detect differences between groups and estimate parameters.

The assumptions 3 and 5 are automatically satisfied since our data set is large consisting of 15,564 colon cancer patients and patients who were lost to follow up have been omitted. Since the predictor for the year of diagnosis (yydx) and the indicator for the year of diagnosis (year8594) are correlated, it is required to remove one of them in order to satisfy assumption 2. Furthermore, assumptions 1 and 4 are investigated later in section 4.3.

Proportional Hazards (PH) assumption

In order to investigate the proportional hazards assumption we illustrate the following methods: the log-log survival curves, Schoenfeld Residual plots and the goodness of fit test. This incorporates both graphical methods along with a statistical analysis using the goodness of fit test.

The log-log survival curve is a transformation of the estimated survival curve which is computed by taking the natural log of a survival probability twice. The hazard function can be rewritten as:

$$S(t|\mathbf{X}) = [S_0(t)]^{e^{\sum_{j=1}^{p}\beta_j X_j}}$$

and then applying the negative of the natural log twice, the expression can be written as:

$$-ln[-lnS(t|\mathbf{X})] = -\sum_{j=1}^{p}\beta_j X_j - ln[-lnS_0(t|\mathbf{X})]$$

Then finally considering two different categories of a covariate corresponding to different individuals $\mathbf{X}$ and $\mathbf{Y}$ and subtracting the log-log curve of the second from the first we get expression:

$$-ln[-lnS(t|\mathbf{X})] = -ln[-lnS(t|\mathbf{Y})] + \sum_{j=1}^{p}\beta_j(X_j - Y_j)$$

This final expression shows that if we use a Cox regression model and plot the log-log survival curves for individuals on the same graph, the two plots should be approximately parallel. The distance between the two curves is the linear expression involving the difference in predictor values, which do not involve time.

The difficulty of this graphical approach is how to decide whether the curves are parallel. This is purely based on observation hence this method assumes the proportional hazards assumption is satisfied unless there is strong evidence of the curves being non parallel. Moreover, it is also important to note if the hazards cross for two or more categories for the explanatory variable of interest then proportional hazards assumption is not met. However even if they do not cross the proportional hazards assumption might still not be met hence we use this graphical method as a baseline to check the PH assumption for categorical variables.

Another graphical method we explored was the plotting of the Schoenfeld Residuals to examine model fit and detect if there we're any outlying covariate values. These residuals represent the difference between the observed covariate and the expected risk set at that particular time. They should be flat and centered at 0 in order to show that the hazard does not have any relationship with time i.e. the hazard is not time dependent in order to satisfy the PH assumption. The idea behind Schoenfeld Residuals is that we first define a partial residual as the difference between the observed value of X_i and its conditional expectation given the risk set R_i and demonstrated that these residuals are independent of time in order to satisfy the PH assumption. Hence, if we represent the residuals ranked by the event time, the plot must not show any pattern for the PH assumption to be satisfied.

Our final method is the the goodness of fit test which goes along with the Schoenfeld residuals plot. The goodness of fit test is a statistical test for the correlation between the Schoenfeld residuals and survival time. A correlation of zero indicates that the model has met the proportional hazards assumption and this is our null hypothesis in this test.

Linearity assumption

It is only essential to check the linearity assumption for the continuous covariates as categorical variables are necessarily linear. This can be done by plotting the Martingale residuals against the continuous covariates in order to detect non-linearity i.e. to assess the functional form of a covariate. The martingale residuals are defined:

$$r_{M_i} = \gamma_i - r_{C_i}$$

There residuals take values between $-\infty$ and unity and residuals for any censored observations, γ_i=0, are negative. These residuals sum to zero and have an expected value of zero. If there is any clear correlation or pattern in the residual plot this indicates to us that the variable might not be linear and therefore might need to be transformed with techniques similar to those used in linear regression. This is achieved by looking at each variable one after another.

2.4 Parametric Methods

The Cox regression model is a useful tool for analyzing survival data because it does not require any specific assumption about the probability distribution of survival times. This means that the baseline hazard function $h_0(t)$ can take on any functional form, making the model very flexible and applicable in a wide range of contexts. In contrast, if a specific probability distribution can be assumed for the data, inferences based on that assumption tend to be more precise, for example, individual survival time. Models that assume a particular probability distribution for survival times are referred to as parametric models.

2.4.1 Exponential model

Exponential models are a type of parametric model used in survival analysis. They assume that the hazard rate $h_0(t)$ is constant over time. In other words, the probability of an event occurring at any given time is proportional to the number of individuals who have not yet experienced the event.

The assumptions for exponential models in survival analysis include:

1. Independence: Each person's survival time in the sample is independent of others.
2. Constant hazard rate: The hazard rate is considered to be constant across time, therefore the probability of an event happening at any given time is the same.
3. Exponential distribution of survival times: It is assumed that the survival times follow an exponential distribution, which has a single parameter that affects the shape of the distribution.
4. The exact survival times of some individuals may be unknown or incomplete due to right-censoring.
5. Homogeneity: It is assumed that all individuals in a population or sample would have the same hazard rate.

The formula for an exponential regression model can be defined as:

$$h_i(t|\boldsymbol{\beta}, \mathbf{X_i}) = \lambda_0 exp(\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi})$$

where $h_i(t|\boldsymbol{\beta}, \mathbf{X_i})$ is the hazard function at time t for ith individual with covariate values $\mathbf{x_i} = (x_{1i}, x_{2i}, ..., x_{pi})$, λ_0 is the baseline hazard rate, and $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ are the regression coefficients associated with each covariate. A common method for estimating the baseline hazard rate in an exponential model is using the maximum likelihood estimation method.

2.4.2 Weibull model

We can extend the exponential model to a more general case known as the Weibull model. Instead of assuming the baseline hazard function is the constant, the Weibull allows the baseline function to vary with time such that:

$$h_0(t) = \lambda \gamma t^{\gamma - 1}$$

for $0 \leq t < \infty$. Particularly, when $\gamma = 1$, it becomes an exponential distribution. When $\gamma \neq 1$, the hazard function should increases or decreases monotonically. Similarly, the hazard function for the ith individual at time t under the Weibull regression model is :

$$h_i(t|\lambda, \gamma, \boldsymbol{\beta}, \mathbf{X_i}) = \lambda \gamma t^{\gamma - 1} exp(\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi})$$

Monotonicity as opposed to homogeneity is the primary difference in assumptions between the exponential model and the Weibull model. This implies that there is a consistent and predictable monotonic relationship between the predictor and response variables *(Collett, 2015).*

3 Results

This section will illustrate the results using the methods described above. All the graphs and calculations are generated using the "Survival" package in R. We began by splitting our data into a 70% to 30% ratio for the training and testing set respectively. The training set is used to train the data to look at possible models and the testing set is used for model validation to evaluate the performance of the model. This split was chosen in order to perform a sufficient analysis of our data whilst also having enough data to gain an accurate prediction of survival time for future patients.

3.1 Analysis of Kaplan-Meier Curves

Using the Kaplan-Meier estimate detailed in the methods section, plots have been produced that highlight the estimated survival rates of patients based on the different independent variables we wish to consider for our final survival time models. By analysing each of the variables in turn, an exploratory analysis has been produced which will help to guide the modelling process later in the report.

Clinical stage at diagnosis: There are visible differences in the Kaplan-Meier curves for the cancer stage variable. Those with a localised cancer diagnosis clearly have a higher chance of survival across the full time period under consideration, with 87.2%, 60.9%, and 41.4% estimated survival rates after 1, 5 and 10 years respectively. Patients with a regional stage diagnosis, where cancer has spread to an adjacent area in the body, have a lower chance of survival than the 'localised' patients, the percentage levels of survivability through time are tabulated below. The survival rates in the case of 'unknown' patients are slightly lower still, where it is unconfirmed as to whether cancer has spread from its area of origin. The patients with the lowest survivability rates have been diagnosed with cancer at the distant stage. The survival rates are 30.7%, 7.5% and 3.04% after 1, 5 and 10 years respectively which can be observed in Table 2.

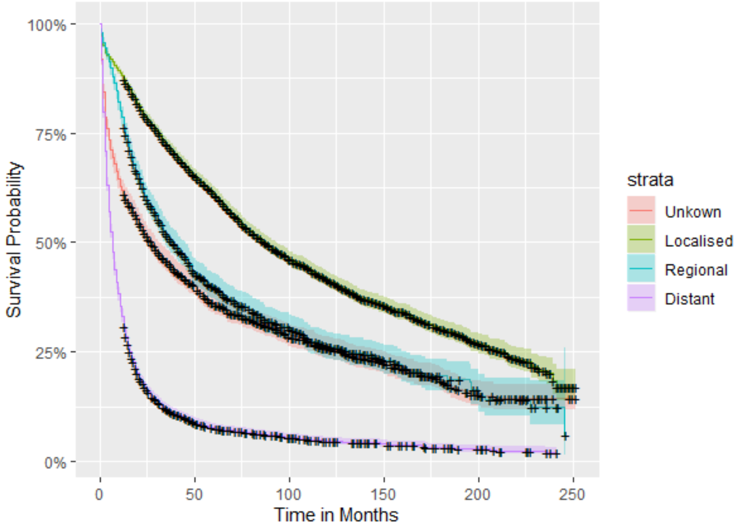


Figure 2: Kaplan-Meier Curves for Cancer Stage

As illustrated in Figure 2, throughout the course of the study, the 95% confidence intervals for the 'regional' and 'unknown' patients overlap significantly. A few different inferences can be made here. The first suggestion is that a large proportion of the 'unknown' patients could have in fact had a regional cancer spread but were not specified this at the time of diagnosis.

One other suggestion is that the unknown patients are a more even mixture of patients from all three stage groups. The overall median survival time of the unknown patients is 26.5 months and when compared with the median of all other patients (27.5 months), it can be noted that the median survivability of the unknown group is similar in value.

Table 2: Estimated survival probabilities for different cancer stages.

	1 year	3 years	5 years	10 years	15 years
Localised	0.872	0.717	0.609	0.414	0.301
Unknown	0.611	0.453	0.358	0.258	0.185
Regional	0.762	0.509	0.397	0.261	0.187
Distant	0.307	0.115	0.075	0.045	0.030

Anatomical subsite of tumour: The 95% confidence intervals for the 4 different subsites appear to overlap for almost all of the time period under consideration, with the exception of the first 2 years. In the initial period after diagnosis, the variation is very small and the confidence intervals do not overlap, illustrating a slight difference in the survival probabilities of patients with different subsites. Evidence for this difference is substantiated by the median survival times of patients; patients with 'Descending and Sigmoid' subsite cancers have a 31.5 month estimated median survival time, whereas subsites 'Coecum and Ascending', 'Transverse' and 'Other and NOS' have medians 27.5, 18.5 and 13.5 months respectively. The Kaplien-Meier plot for the subsite can be observed in Figure 3.

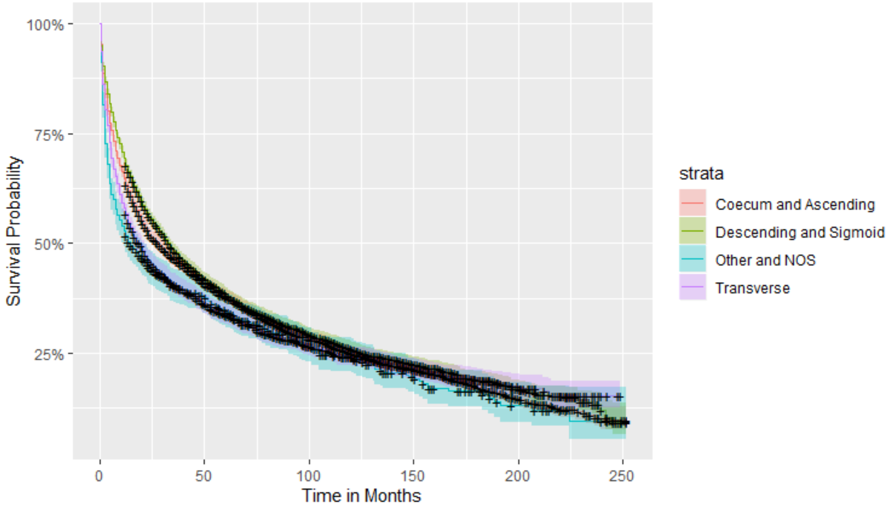
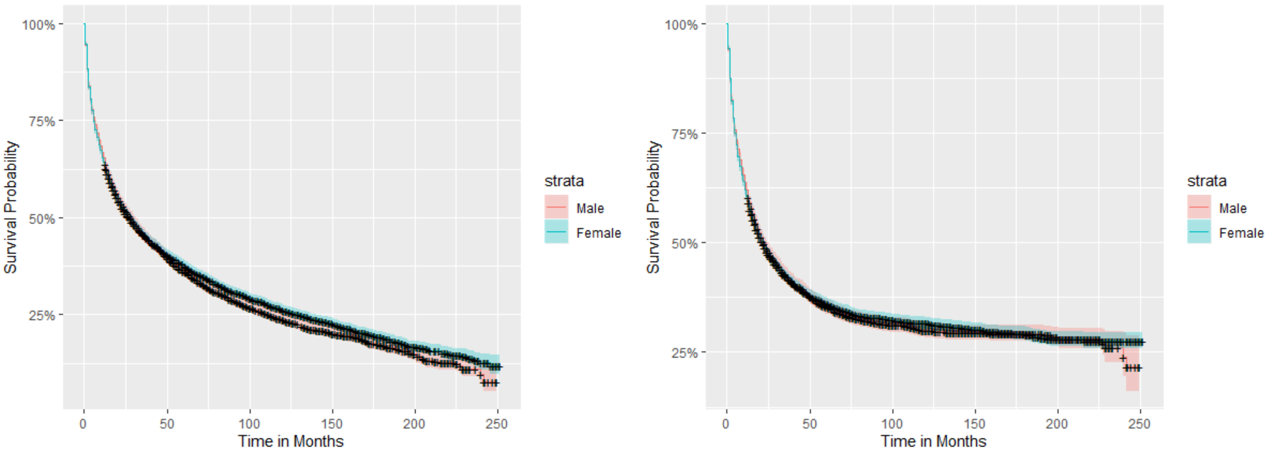


Figure 3: Kaplan-Meier Curves for Cancer Subsite

Sex of patient: The sex of the diagnosed patient does not seem to have any discernible association with the estimated survival probability. In the initial stages after the diagnosis, the estimated survival times are extremely similar, however, after around 75 months, slight differences start to emerge with males having a smaller survival probability. As in the case of other factors, the variation of these estimated probabilities increases with time. Considerations must be made here to the fact that the life expectancies of males and females differ globally, more specifically, women have a longer life expectancy than men in Europe where the study was conducted *(Clark and Peck 2012)*. To explore this, the patients who had died 'with cancer', were temporarily removed from the dataset to see the effect on the Kaplan-Meier curve when only those who died as a direct consequence of colon cancer were considered. On the new curve, there is no real noticeable difference in the survival probabilities of patients throughout the whole time period.



(a) All patients

(b) Died of Cancer

Figure 4: Life expectancy bias for Sex

Respiratory Complications: There does not seem to be any discernible visible association of respiratory complication status with the estimated survival rates of patients. As expected, the survival time variance of those with complications is much higher as these patients only account for a small proportion of the records in the dataset (3%). The blue confidence interval represents the confidence interval for patients in the N/A category. This is very large because there are few patients in this category, and therefore the blue curve can be interpreted as unreliable.

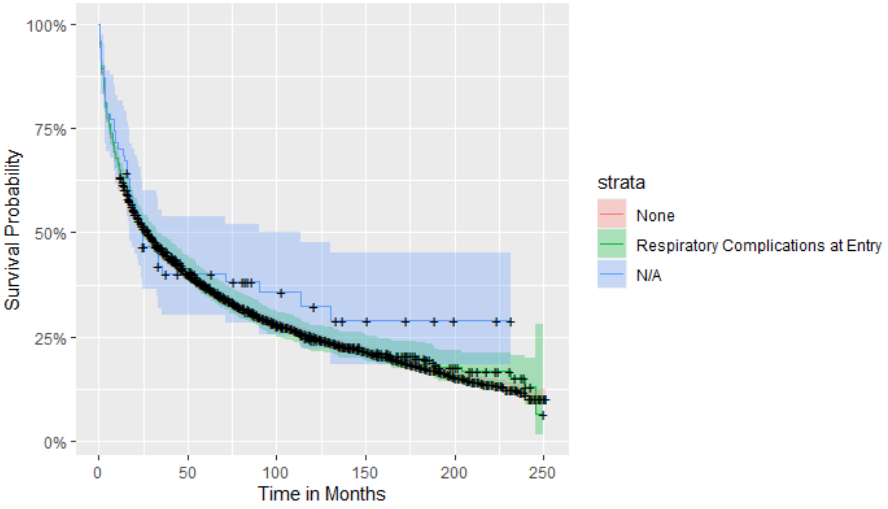
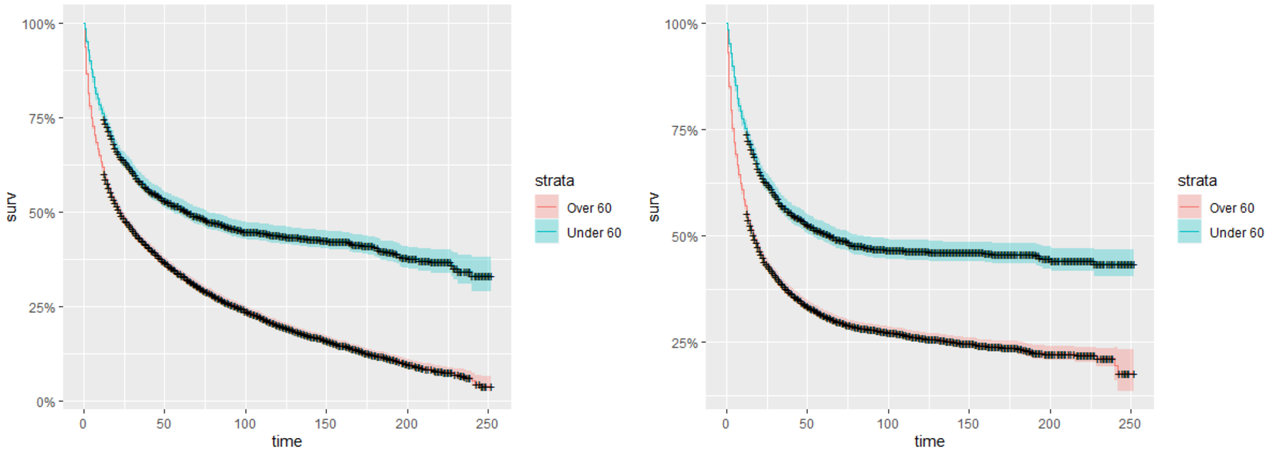


Figure 5: Kaplan Meier-Curves for Respiratory Condition

Age at diagnosis: To interpret and explore the association of a patient's age with their estimated survival probability, patients have been grouped into under 60 and over 60. In the initial plot, which includes all deaths both 'of 'and 'with' cancer, a clearly discernable pattern is shown, patients under 60 have a much higher estimated survival probability than patients over 60. This disparity increases with the time since diagnosis, as those diagnosed in old age become increasingly likely to die from other causes as time increases. As with the sex variable, it is useful to explore the curves that only include those who died as a direct consequence of the cancer to determine the significance of age. This is because those diagnosed over the age of 60 are much more likely to die of other causes such as old age. As seen in Figure 5, there is still a large disparity in the survival probabilities of those diagnosed over the age of 60 and those diagnosed under the age of 60, although the survival probability of those diagnosed at an old age does increase slightly when eliminating deaths by natural causes.



(a) All patients

(b) Died of Cancer

Figure 6: Kaplan-Meier Curves for Age

Year of Diagnosis: Figure 7 compares the Kaplan-Meier curves for those diagnosed in the years 1975-84 with those diagnosed in the years 1985-94. Interestingly, the curve for 1985-94 terminates after 130 months. This is because the patients in this category were only followed for a maximum of 130 months, compared to the 250 for the former. Despite the confidence intervals not overlapping, the curves up until this point follow a close path. From the curves, it is not conclusive as to whether this variable has a significant effect on survival time. However, we suspect that technological advancements across the time period might make the year of diagnosis significant, and we expect to include this in the modelling process.

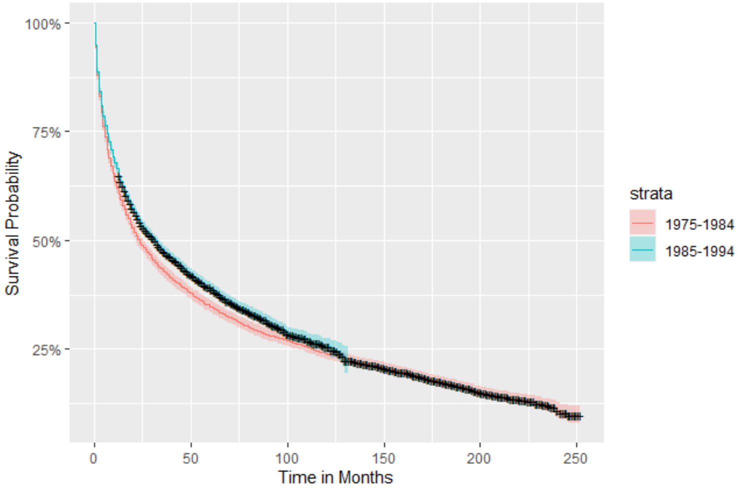


Figure 7: Kaplan Meier-Curves for Year Diagnosed

3.2 Log-rank test

The difference in survival time between different stages of cancer was compared by performing a log-rank test. The null hypothesis states that the survival time for all the stages is the same, whereas the alternative hypothesis states that there is at least one stage that has affected the survival time. The p-value is 10^{-16}, which means that there is overwhelming evidence to reject H_0. Therefore, there is strong evidence to say that the stage has a significant effect on survival time for colon cancer patients. This result could also be corroborated by the survival curve for different stages in Figure 2, where the survival probability is distinct for different stages over time.

Similarly, log-rank tests were performed for the factor variables including the anatomical subsite, respiratory complications at entry and sex of a patient.

For the log-rank test for anatomical subsite, the null hypothesis states that the survival time for all the subsites is the same, whereas the alternative hypothesis states that there is at least one subsite in which the patient's survival time is different from others. The p-value is 10^{-9}, which means that there is strong evidence to reject H_0. Therefore, there is strong evidence to say that different subsites for colon cancer patients have different survival times. The survival curves in Figure 3 seem to be similar to each other, however, there is an obvious difference when considering the 'Descending and Sigmoid' and 'Other and NOS' curves in Figure 3.

For the log-rank test for sex, the null hypothesis is there is no difference in survival time between male and female colon cancer patients, and the alternative hypothesis is there are differences in survival time between male and female colon cancer patients. The p-value calculated is 0.2, which means that there is no evidence to reject the null hypothesis. Therefore, there is no evidence to say that sex has an influence on survival time for colon cancer patients. The result could be corroborated by the survival curve for males and females as shown in Figure 4, the two curves for males and females are very similar to each other.

The null hypothesis of the log-rank test for respiratory complications at entry is whether the patients have respiratory complications at entry not affect their survival time, and the alternative hypothesis is there are differences in survival time. The p-value is 0.6 which means that there is no evidence to reject the null hypothesis, hence there is no evidence to say that respiratory complications at entry affect the survival time of colon patients.

The results of the log-rank test only look at each categorical variable separately, however, when analysing these variables along with other variables in the later section, the result might be different.

3.3 Cox Regression

3.3.1 Model Selection

We began by using a multivariate Cox regression model of the full model in order to see the significance of each variable with respect to survival time. The null hypothesis for this model is that there is no association between the survival time and the variable of interest.

From the exploratory analysis, we noticed that the correlation of the two variables yydx and year8594 was 0.861, and the p-value of their correlation test was 10^{-16}, which suggests that they are highly correlated. Therefore, we will only consider the variable yydx in the Cox regression model. Then we proceeded to look at the model selection for the Cox regression model and discovered using yydx as opposed to year8594 decreases the values of AIC and BIC.

Full model

We began by applying multivariate cox regression to observe the full model. The summarised results for each predictor variable are detailed below.

Age: The hazard ratio is 1.038 (and the regression coefficient is positive) therefore the hazard is higher for older patients hence age is positively associated with death and so is negatively associated with the length of survival. The p-value gives overwhelming evidence to reject the null hypothesis, hence we conclude that age has a significant effect on the survival time of patients with colon cancer.

Sex: The hazard ratio is 0.849 (and the regression coefficient is negative) hence hazard decreases for patients that are female and hence being female is associated with good prognostic. At a given instance in time, we find that someone who is male is 1.17 times as likely to die as someone who is female adjusting for other factors. The p-value shows overwhelming evidence to reject the null hypothesis and so sex is also a significant variable.

Stage: The hazard ratio for the distant stage is 2.872 (and the regression coefficient is positive) showing that the hazard is higher for patients with a distant stages compared to other stages as patients at a localised stage and a regional stage have a hazard ratio of 0.555 and 0.957 respectively. As the hazard ratio is smaller than one for these stages this shows that the hazard decreases at these stages. The p-value again shows overwhelming evidence at each level to reject the null hypothesis and so we have strong evidence to show stage is a significant variable

Mmdx and Respiration: The p-values obtained show we don't have enough evidence to reject the null hypothesis and hence this indicates that these variables are not strongly related to the survival rates of colon cancer patients.

Yydx: The hazard ratio is less than one (and the regression coefficient is negative) meaning that the hazard decreases for patients diagnosed in later years. This is intuitive, since through time technological advancements occur and a wider range of information is accessible to healthcare professionals. Again, the p-value gives overwhelming evidence to reject the null hypothesis and so yydx is a significant variable.

Subsite: The hazard ratio is 0.983 (and the regression coefficient is positive) showing that the hazard is higher for patients with a higher subsite factor, in other words, patients with a transverse subsite have higher hazard than the other patients at the other subsites. The p-value again shows strong evidence to reject the null hypothesis and so we have evidence to show subsite is a significant variable.

Variable Selection Using AIC and BIC

The Akaike's information criterion (AIC) and Bayesian Information Criterion (BIC) methods are used in the variable selection procedures. The values of AIC and BIC increase as extraneous variables are included in the model, therefore, smaller values of AIC and BIC indicate better models.

The results of the Kaplan-Meier curves and the log-rank test have shown that the factor variables: stage and subsite had a significant effect on survival time. Therefore, they should be kept in our model when considering different models.

Based on the findings in the above analysis, all of the possible Cox regression models were fitted and compared based on the values of AIC and BIC of these models. From Table 3 below we can see that the variables that were the most significant are: subsite, stage, yydx, age and sex. This is because the model including these variables has the lowest AIC and BIC values.

Table 3: Values of AIC and BIC for Cox regression models using data from the training set

Variables	AIC	BIC
subsite + stage	129306.7	129320.5
subsite + stage + yydx	129273.4	129294.2
subsite + stage + age	128044.8	128065.7
subsite + stage + sex	129308.5	129329.3
subsite + stage + resp	129308.8	129336.5
subsite + stage + yydx + age	127969.9	127997.6
subsite + stage + yydx + sex	129275	129302.7
subsite + stage + yydx + resp	129275.7	129310.4
subsite + stage + age + sex	128011.1	128038.9
subsite + stage + age + resp	128048.2	128082.9
subsite + stage + sex + resp	129310.5	129345.2
subsite + stage + age + sex + resp	128014.1	128055.8
subsite + stage + yydx + sex + resp	129277.2	129318.8
subsite + stage + yydx + age + resp	127973.5	128015.1
subsite + stage + yydx + age + sex	127929.9	127964.6
subsite + stage + yydx +age + sex + resp	127933.2	127981.8

3.3.2 Proportional Hazards Assumption

Log-log curves

As seen in section 2.3, the log-log curves are a graphical technique that allows us to observe whether the proportional hazards assumption is met. The curves for each level of the categorical variables should be parallel and they should not cross at any point.

The log-log curves for sex, stage and subsite with log-log survival time against survival time in months can be seen in Figure 8 and 9 below. From the plots, it is clear to observe that the curves cross for each of the levels in the categorical variables and so the PH assumption is violated for all the categorical variables of interest. The log-log curves of resp and year8594 can be seen in the Appendix (Figure 20 and 21) which also shows a violation of the PH assumption for both variables.

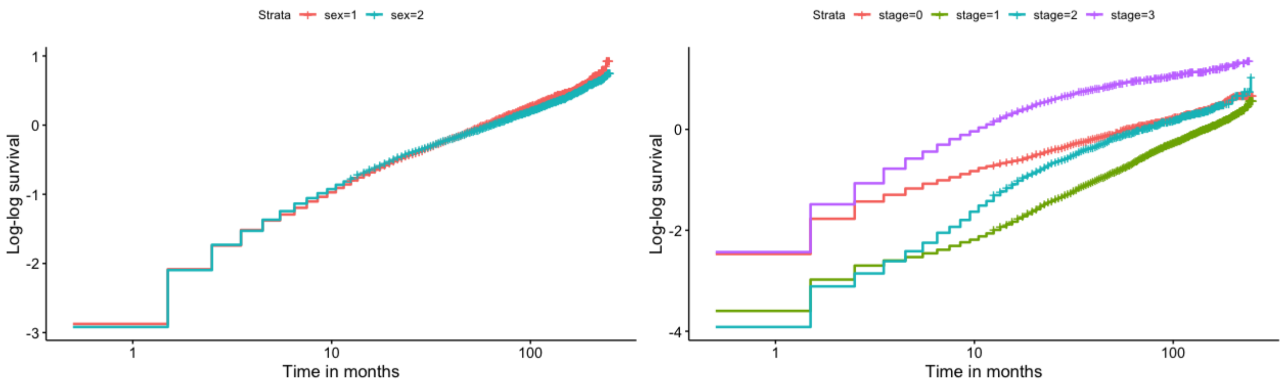


Figure 8: Log-log curve for sex and stage

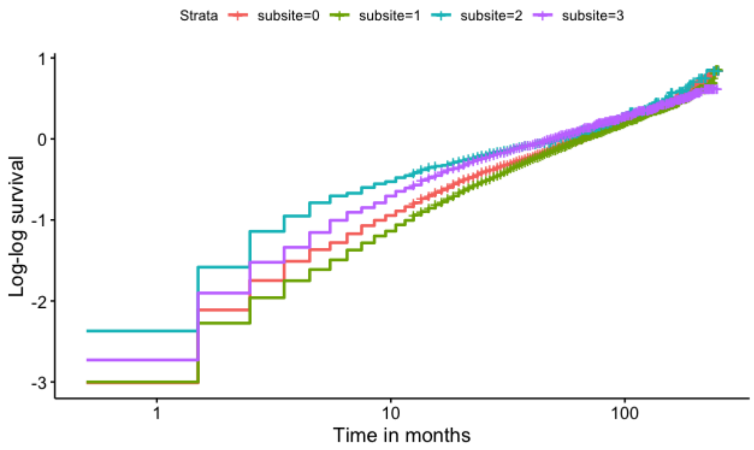


Figure 9: Log-log curve for subsite

Schoenfeld Residuals and goodness of fit tests

As seen in section 2.3, the Scheonfeld Residual plots are used to assess whether the hazards are proportional to time for each of the variables of interest. Therefore, if the residual plots do show a correlation to time this implies that the PH assumption is violated. The goodness of fit test as mentioned in the methods will show us statistically whether there is a correlation between the Schoenfeld residuals and survival time.

We can observe below, Figure 10 shows that for the variables, age and yydx the residuals are horizontal and randomly distributed at approximately 0 hence we find these variables do not violate the PH assumption by observation.

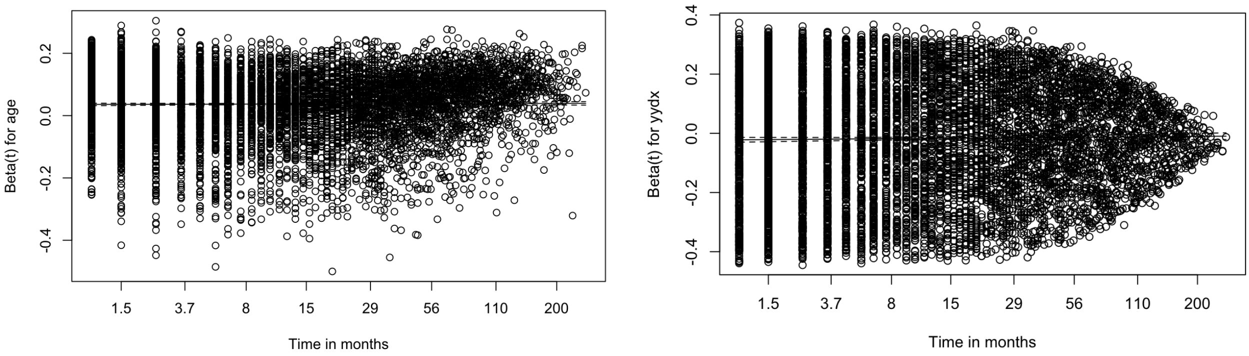


Figure 10: Schoenfeld residuals plot of age and yydx

The residual plot of the stage and subsite can be seen in Figure 11. The variable stage shows a clear violation of the proportional hazards assumption as the residuals show a clear correlation with time and are not distributed about 0. Similarly, for subsite we see the same results as stage, however it's not as obvious from observation as the stage variable. Consequently we will have to investigate this further when using our final method to check the proportional hazards assumption; the goodness of fit tests.

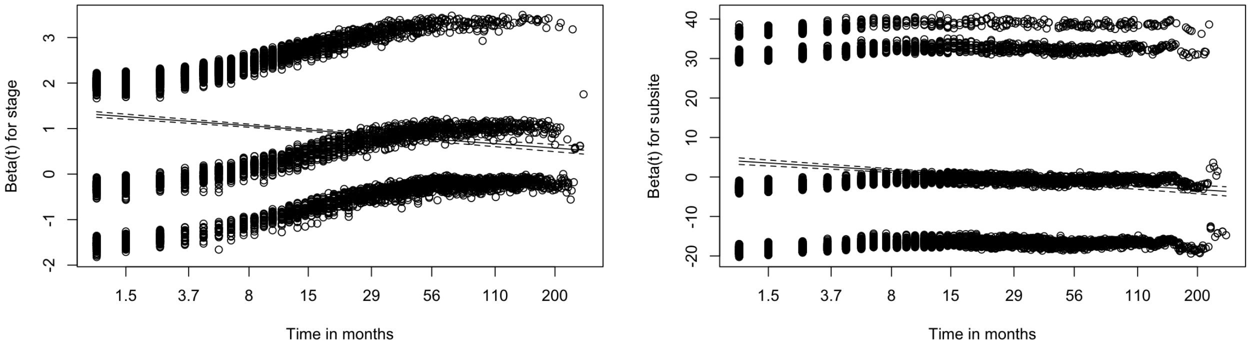


Figure 11: Schoenfeld residuals plot of stage and subsite

Finally, from Figure 12 we observe the Schoenfeld residuals plot for sex which does not show any clear correlation with time and are approximately distributed about 0, however again we need to check the goodness of fit tests in order to to see whether our results are valid from this graphical method statistically.

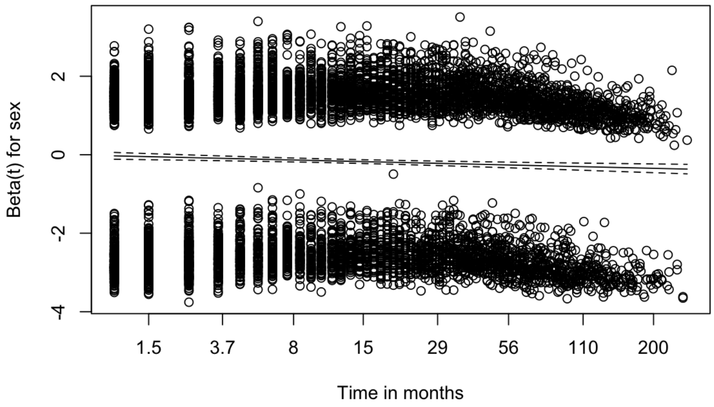


Figure 12: Schoenfeld residuals plot of sex

The Schoenfeld residuals again only show a graphical interpretation and it can sometimes be ambiguous as to how to interpret the residuals hence we look at the goodness of fit tests to give us the statistics as to which variable is violating the proportional hazards assumption.

The goodness of fit test for each of the variables can be seen in Table 4.

Table 4: Goodness of Fit test

Explanatory Variable	p-value
Age	0.46421
Sex	0.00023
Stage	$< 2e - 16$
Year of Diagnosis(yydx)	0.02736
Subsite	$< 2e - 16$

The result from Table 4 confirms the findings in the Schoenfeld residual plots for age and yydx. Although yydx has a smaller p-value the null hypothesis is satisfied at the 1 % value and the Schoenfeld residuals plot for yydx also shows no correlation with time. Therefore, it is interpreted that yydx does satisfy the PH assumption which can be verified after we adjust for the variables which do not satisfy the assumption. The Table shows overwhelming evidence that sex, stage and subsite violate the proportional hazards assumption. This was clear from the Schoenfeld residuals plot for stage and subsite. However, for sex we now find that it violates the proportional hazards assumption through this test as there is overwhelming evidence to reject the null hypothesis. This is why it's important to use the goodness of fit test along with the residual plots as it's a clear statistical way to see if there is any correlation between time and the Schoenfeld residuals. Now it is essential we consider methods to deal with this violation such as the stratification of variables.

3.3.3 Stratification

From Table 4, it is apparent that three of the variables: subsite, stage and sex, violate the PH assumption at the 1% level. As a result of the conclusions made previously regarding the PH assumption, a stratified Cox regression model is set up that stratifies the three variables, subsite, stage and sex.

As we need to stratify three variables, a single new categorical variable Z* should be established. The categories of Z* represent the combinations of categories of the three variables. The subsite and stage variable has four categories by definition. The sex variable has two categories by definition. Thus, the number of categories for our Z* variable is the product of 4, 4 and 2, implying that Z* has 32 categories.

The no-interaction model is defined by the hazard function below:

$$h_g(t, \mathbf{X}) = h_{0g}(t) exp[\beta_1 yydx + \beta_2 age], g = 1, 2, ..., 32$$

To evaluate whether the no-interaction model is appropriate, an interaction model that allows different regression coefficients for different strata is also detailed below:

$$h_g(t, \mathbf{X}) = h_{0g}(t) exp[\beta_{1g} yydx + \beta_{2g} age], g = 1, 2, ..., 32$$

An alternative version of this interaction model that involves product terms is shown here. This version uses 31 dummy variables denoted as Z_1^*,Z_2^* up through Z_{31}^* to distinguish the 32 categories of the stratification variable Z*. The model contains the main effects of yydx and age plus interaction terms involving products of each of the 31 dummy variables with each of the two predictors *(Kleinbaum and Klein 1996).*

$$\begin{aligned} h_g(t, \mathbf{X}) = h_{0g}(t) exp\{\beta_1 yydx + \beta_2 age + \beta_{1,1}(Z_1^* \times yydx) + ... + \beta_{1,31}(Z_{31}^* \times yydx) \\ + \beta_{2,1}(Z_1^* \times age) + ... + \beta_{2,31}(Z_{31}^* \times age)\}, g = 1, 2, ..., 32 \end{aligned}$$

Another version of interaction model: Replace $Z_1^*, ..., Z_{31}^*$ by

Z_1^*=Descending and sigmoid(binary)
Z_2^*=Other and NOS(binary)
Z_3^*=Transverse(binary)
Z_4^*=Localised(binary)
Z_5^*=Regional(binary)
Z_6^*=Distant(binary)
Z_7^*=Female(binary)
$Z_8^* = Z_1^* \times Z_4^*$
$Z_9^* = Z_2^* \times Z_4^*$
......
$Z_{31}^* = Z_3^* \times Z_6^* \times Z_7^*$

32 possible combinations of Z_1^* to Z_{31}^*:
$g = 1 : Z_1^* = Z_2^* = Z_3^* = Z_4^* = Z_5^* = Z_6^* = Z_7^* = 0$
$g = 2 : Z_1^* = 1, Z_2^* = Z_3^* = Z_4^* = Z_5^* = Z_6^* = Z_7^* = 0$
$g = 3 : Z_2^* = 1, Z_1^* = Z_3^* = Z_4^* = Z_5^* = Z_6^* = Z_7^* = 0$
......
$g = 32 : Z_3^* = Z_6^* = Z_7^* = 1, Z_1^* = Z_2^* = Z_4^* = Z_5^* = 0$

Our null hypothesis, H_0, is such that a no-interaction model acceptable, i.e., yydx: $\beta_{1,1} = \beta_{1,2} = ... = \beta_{1,31} = 0$ and age: $\beta_{2,1} = \beta_{2,2} = ... = \beta_{2,31} = 0$
The null hypothesis tests whether the no-interaction model is acceptable. As the null hypothesis involves 62 coefficients, the degrees of freedom of the LR chi-square statistic is 62, which means that $LR \sim \chi^2_{62}$ under H_0: no interaction.

$LR = -2lnL_R - (2lnL_F)$, where lnL_R is the log-likelihood statistic for the reduced(no-interaction) model and lnL_F is the log-likelihood statistic for the full(interaction) model. Hence, we calculate that the LR statistic which is 424.7267. Therefore, we get that the p-value is approximately 0, which gives a significant result at the 1% level. This indicates that the no-interaction model is not acceptable and the interaction model is preferred.
After stratification we look again at the goodness of fit tests and find that age and yydx now satisfy the proportional hazards assumption with p-values 0.083 and 0.159 respectively.

3.3.4 Linearity Assumption

Martingale residuals

The plot below, Figure 13 highlights the Martingale residuals of age before and after we add a polynomial term. By using LOESS (Locally weighted scatterplot smoothing) regression to plot a smooth curve through the residuals, it becomes evident that age might not be a linear term as the residuals show a correlation and do not have an expected value of zero. After a squared term was added to age, the curve becomes horizontal about 0, this is an improvement upon the previous residual plot indicating that a quadratic term for age provides a better fit for our dataset.

Higher degree polynomials such as cubic terms were also considered for age. However, after conducting anova tests with a cubic term there was no evidence at the 5% level to reject our model with a quadratic age term, therefore we decided to model age with a quadratic term.

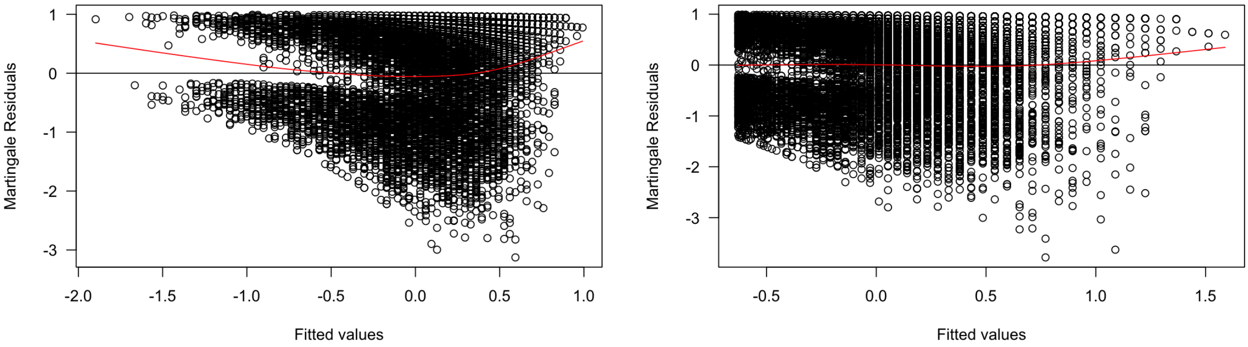


Figure 13: Martingale residuals plot of age and polynomial age transformation

Additionally, the Martingale residuals for yydx has a horizontal LOESS plot through the residuals therefore this indicates no transformation of yydx is needed and yydx can be kept as a linear term which can be seen in the Appendix at Figure 22.

3.3.5 Estimating the baseline hazard and survivor function

Consequently to work out the survival time we need to first be able to estimate the baseline hazard. As we have seen after stratification the baseline hazard is different for each strata. To find the distribution our baseline hazard follows maximum likelihood estimation is used. This method includes fitting different parametric distributions such as the Weibull, log-normal and exponential and then assessing these fits using AIC and BIC methods which we used previously in section 4.3.1. The results of this method can be seen in Table 5. The log-likelihood value of the distribution of interest is a goodness of fit test to assess the fit of a model and the values range from $-\infty$ to ∞.

Table 5: Maximum Likelihood estimation

Parametric Distribution	Log-likelihood values	AIC	BIC
Weibull	-174.8069	353.6138	360.1392
Exponential	-199.7752	401.5504	404.8131
Log-normal	-200.6283	405.2566	411.782

This allows us to validate that the Weibull distribution is the most appropriate for modelling the baseline hazard. This report will just show the stratification for patients that are female, at a localised stage of cancer with a descending and sigmoid subsite of colon cancer. As each strata has a seperate baseline function and we have many strata, we will convey our prediction using this particular strata. The specific values of the parameters can be found in appendix in Table 13 and the baseline function definition of a Weibull model can be found in section 2.4.2.

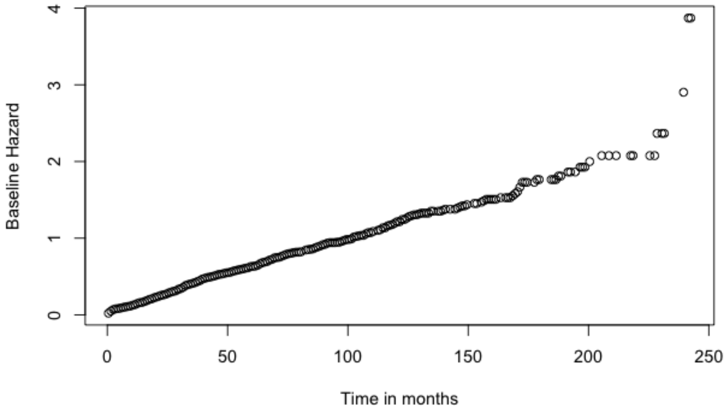
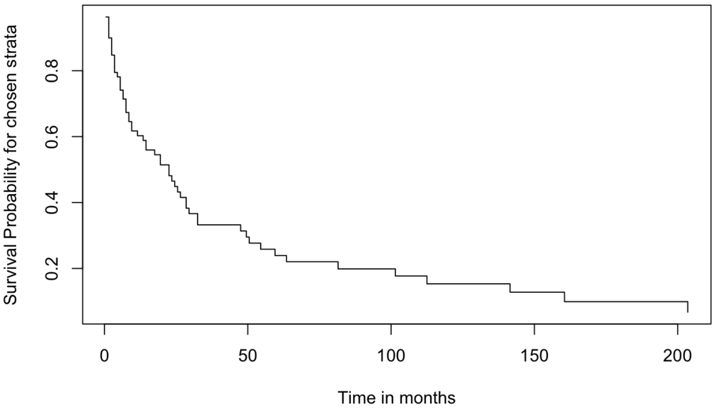


Figure 14: Baseline hazard plot of the first strata

For each strata we have a different survivor function, we first began by plotting the survivor function through calculating the cumulative hazard function and using the definition which can be seen in section 2.1.1. The graph of the survival curve for the strata of interest is below.



Hence modelling our stratified model when the baseline hazard and survivor function follows a Weibull distribution allows us to predict the survival time which we can see in Figure 15.

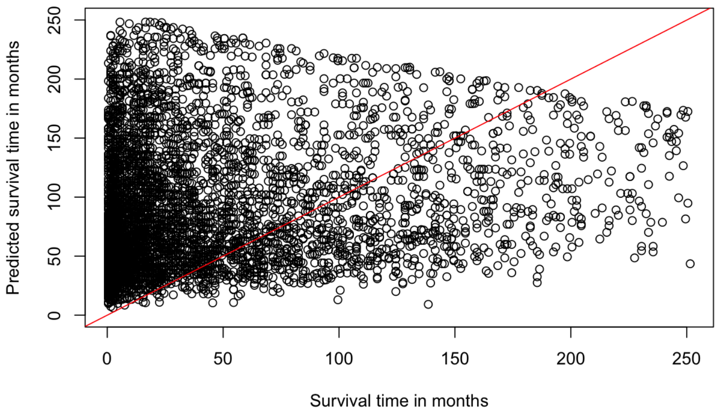


Figure 15: Stratified Cox regression model predictions

The mean squared error for this model is relatively small, with values seen in Table 6 below.

Table 6: Mean Squared Error for cox model

Model	MSE by Months	MSE by Years
age + I(age^2) + yydx + strata(subsite, stage, sex)	6918.44	35.48

Now we want to look at this for an individual. An individual in the data set was chosen to see how the estimated survival time compares with the actual survival time in our results. We chose at random, a 77-year-old female, at a localised stage of cancer with a descending and sigmoid subsite of colon cancer. From our model predictions we find that she has 46.235 months (3.85 years) left of survival from the time of diagnosis. The individual chosen had an actual 46.5 months (3.5 years) of survival time. The range of this is only 0.25 years indicating our model is a good fit for the data.

Furthermore, this survival curve was used to estimate the probability of an individual surviving during different time periods which can be seen in Table 7 below.

Table 7: Median and mean survival time for Cox regression model

Years of survival	Median survival time	Mean survival time
5 years	0.521	0.545
10 years	0.478	0.565
15 years	0.446	0.565
20 years	0.449	0.464

The mean survival time increases from 5 years to 10 years implying that the survival curve decreases quicker in the first 5 years than the future 10 years. This means that the risk of dying is higher in the first 5 years and then decreases in the future 10 years. This is a common observation for cancer as it implies that the treatment, for example chemotherapy or radiation therapy, used to treat colon cancer might have killed cancer cells and decreased the risk of dying in the 10 year survival period.

The mean survival time decreases from 15 years to 20 years. This implies that the survival curve starts to decrease quicker in the 20 years than in 15 years time period. This might be because the colon cancer might have reoccurred and so increases the risk of dying in future years hence decreasing the survival probability of an individual. This is a frequent occurrence in cancer research.

We observed similar results for each different strata hence we only show this example strata in our results.

3.4 Exponential model

3.4.1 Model Selection

The exponential model was fitted using the same training data set as above. Since yydx and year8594 were highly correlated, two versions of the full model were built using year8594 and yydx respectively. The summary output from R indicated that resp was not significant in both models and therefore was removed from the model. For the model with yydx, yydx was not significant with a p-value of 0.12413, therefore it should be removed.

The AIC and BIC method was also applied to model selection.

Table 8: Values of AIC and BIC for exponential models using data from the training set

Variables	AIC	BIC
subsite + stage + sex + yydx + age + resp	73021.08	73079.42
subsite + stage + sex + year8594 + age +resp	73019.11	73077.45
subsite + stage + sex +year8594 + age	73016.68	73060.43

3.4.2 Assumption check

The assumptions of the exponential model were discussed in section 4.4.1. First, we use a residual plot to check if the constant hazard rate is satisfied. A residual plot is a graphical tool used to check if the residuals are randomly distributed around zero, which is an indicator of a constant hazard rate.

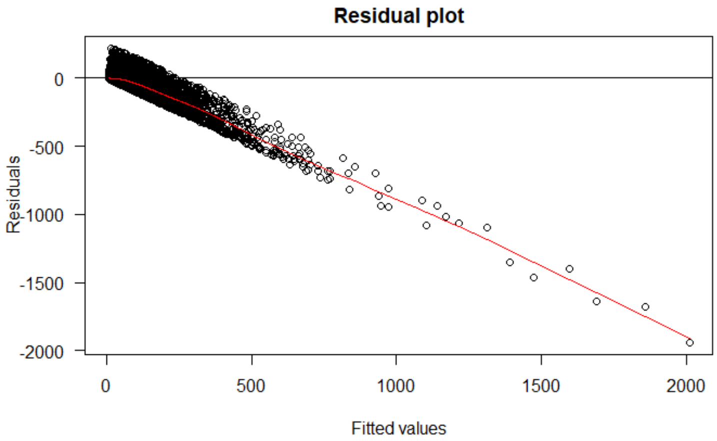


Figure 16: Residual plot of the exponential model

Figure 16 shows that the residual of the exponential model is not randomly distributed around zero, it suggests that the assumption of constant hazard rate over time is violated. Therefore, alternative models such as the Weibull models should be considered.

3.5 Weibull model

3.5.1 Model Selection

Full Model

The functions in the “SurvRegCensCov” package were used to perform Weibull regression, which returned coefficients(β_i), standard error, p-value, hazard ratio, model’s scale and so on. As before, we only used one of the yydx and year8594 variables in the full model.

We first looked at the summary of the full model [sex + stage + subsite + yydx + age + resp]. It showed that the p-value of resp is 0.85, which indicated that there was no strong evidence that the coefficient for that predictor was different from zero. Then we compared Model 1 [sex + stage + subsite + yydx + age] and Model 2 [sex + stage + subsite + year8594 + age], which suggested that “yydx ” was better due to the smaller p-value.

Moreover, the summary of Model 1 demonstrated that the two categories with p-value larger than 0.1 were stage[Regional] (0.4796) and subsite[Descending and sigmoid](0.1871). Since the baseline function was based on stage[Unknown] and subsite[Coecum and ascending] respectively, we need to combine them together: (Regional + Unknown), (Coecum + Descending). Additionally, after first combining the categories, the summary suggested that subsite[Transverse] and subsite[Other and NOS] should be bounded as (Other + Transverse) due to the p-value equalling 0.851. This model was defined as Model 3, whose scale was 1.37 and other parameters are given in Table 9:

Table 9: The Summary Result of Model 3

Parameters/Variables	Estimated coef	Hazard ratio	p-value
λ	6.130013e+07	/	/
γ	7.296134e-01	/	/
sex_Female ($X_1 = 1$; else = 0)	-1.677928e-01	0.8455290	1.4e-12
stage_Distant ($X_2 = 1$; else = 0)	1.714155e+00	5.9837	< 2e-16
stage_Unknown+Regional ($X_3 = 1$; else = 0)	5.906887e-01	1.8052312	< 2e-16
subsite_Other+Transverse ($X_4 = 1$; else = 0)	1.083293e-01	1.1144146	5.1e-05
age (X_5 cts)	3.841186e-02	1.0391591	< 2e-16
yydx (X_6 cts)	-1.216735e-02	0.9879064	3.0e-08

Thus, the complete hazard function of ith individual at time t under Model 3 is shown below:

$$h_i(t) = 6.13\times10^7\times0.7296t^{-1.7296}exp(-0.1677x_{1i}+1.7142x_{2i}+0.5906x_{3i}+0.1083x_{4i}+0.0384x_{5i}-0.0122x_{6i})$$

Further Selection: AIC and graphical goodness-of-fit test

Having devised a model whose predicted variables were highly related, some variables were removed to consider the over-fitting problem. We used the AIC method shown in Table 10:

Table 10: Weibull Model with AIC Selection

	Model with minimum AIC	AIC value
1 predictor	stage_new	71441.68
2 predictors	stage_new + age	70173.16
3 predictors	stage_new + age + sex	70130.34
4 predictors	stage_new + age + sex + yydx	70101.8
5 predictors	stage_new + age + sex + yydx +subsite_new1	70087.71

This was unsuccessful since the model with the smallest AIC was stage_new + age + sex + yydx +subsite_new where stage and subsite had already been combined. Thus, we had to look at the graphical goodness-of-fit test. Figure 17 shows the most fitted model with four variables, stage_new + age + sex + subsite_new [Model 4].

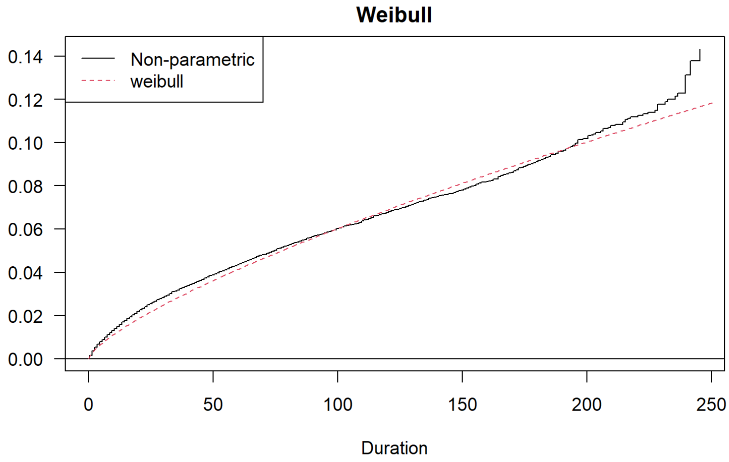


Figure 17: Graphical goodness-of-fit test of stage+ age+sex+subsite

Thus, the complete hazard function of ith individual at time t under Model 4 is

$$h_i(t) = 0.0020 \times 0.7361t^{-1.7361} exp(-0.1580x_{1i} + 1.7172x_{2i} + 0.5876x_{3i} + 0.1059x_{4i} + 0.0377x_{5i})$$

where,
Sex: Male[$x_1 = 0$], Female[$x_1 = 1$];
Stage: Distant[$x_2 = 1, x_3 = 0$], Unknown or Regional[$x_2 = 0, x_3 = 1$],Localised[$x_2 = 0, x_3 = 0$] ;
Subsite: Coecum or Descending[$x_4 = 0$], Other or Transvers[$x_4 = 1$];
Age: X_5, which is a continuous variable.

3.5.2 Assumption check

All the assumptions are the same as before except for Monotonicity. According to the article *(Zhang, 2016)*, we drew plots of log[survival time] versus log[–log(KM)] to check whether the lines were linear and parallel for categorical variables. For continuous variables age and yydx, we stratified them by the threshold value 65 age and 1985 year.

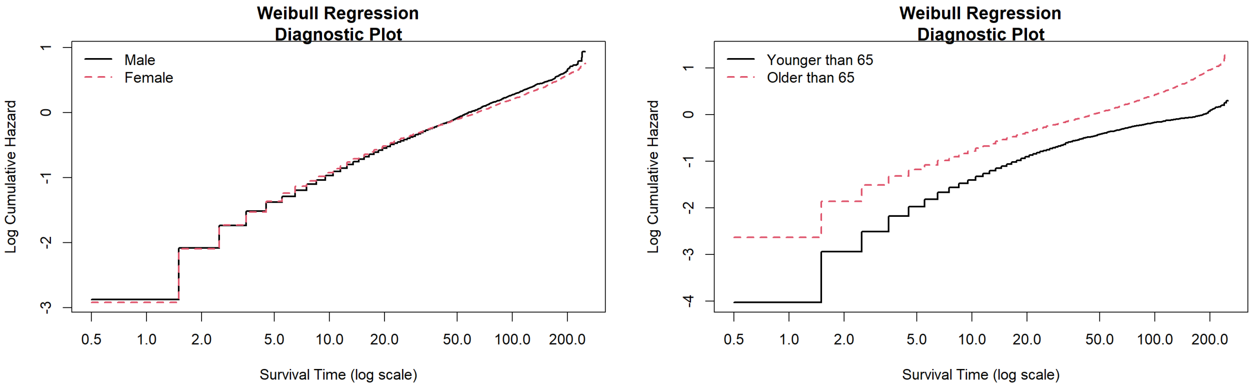
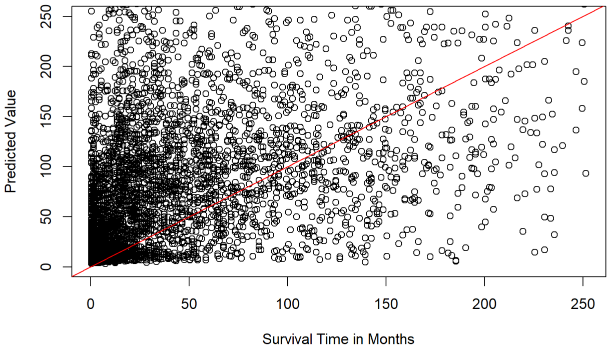


Figure 18: log-log plot of sex and age

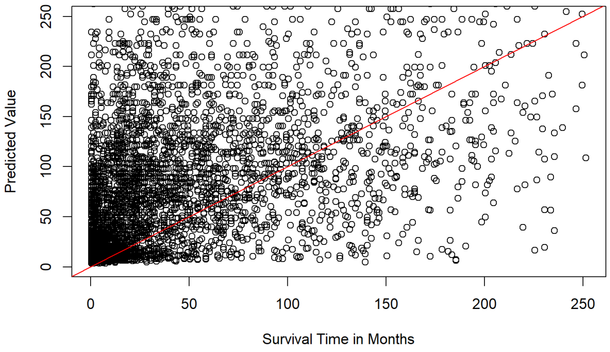
According to Figure 18, all the variables satisfied the assumption. Other plots are attached in the Appendix (Figure 23, 24 and 25).

3.5.3 Prediction

In the prediction part, we first combined categories of variables stage and subsite. Then we used the function "predict" and drew the scatter plot of the predicted value against the actual value:



(a) Model 3: sex + stage + subsite + age + yydx



(b) Model 4: sex + stage + subsite + age

It was hard to say which model performed better in this dataset, but both of them are not perfectly fit. We calculated Mean Squared Error (MSE) for these two models:

Table 11: Mean Squared Error For Two Models

	MSE by Months	MSE by Years
Model 3: sex + stage + subsite + age + yydx	20405.73	141.2754
Model 4: sex + stage + subsite + age	17523.93	121.273

Here, MSE by Years means using surv_yy to compare with the predicted value divided by 12.

Model 4 performed better in this dataset. We can change different dataset to test these two models to decide which fits well.
Here are some examples:

Table 12: Prediction examples using Weibull regression model

Patient's ID	Surv_yy	Model 3_Pred	Inference of Model 3	Model 4_Pred	Inference of Model 4
77	8.5	8.0376093	[7.4171967, 8.6580219]	9.0597004	[8.4727908, 9.6466099]
3502	0.5	0.4498012	[0.4035934, 0.4960090]	0.5315825	[0.48717, 0.5759878]
10345	6.5	94.44936	[81.16278, 107.73595]	82.15665	[71.48752, 92.82578]

In the third row of the example, predictive intervals are far away from the true value. This situation happened when the patients' age was very young, for example, 34 years old. It needs further research in the future.

4 Conclusion

In conclusion, the variables that we considered to have a significant impact on survival time based on the Kaplan-Meier curves were the anatomical subsite of tumour and the stage of cancer at diagnosis. These conclusions were verified using the log-rank test, which outputted p-values suggesting similar findings as expected. This gave us an initial insight into which variables might be useful to include in our statistical models, in order to obtain better predictive power.

Having explored the significance of each variable in our model selection process, and checking the assumptions of the model, we proceeded to stratify variables which violated the PH assumption. Hence, the final Cox regression model is defined below as:

$$h_g(t, X) = h_{0g}(t)exp\{\beta_{1g}yydx + \beta_{2g}age + \beta_{3g}age^2\}, g = 1, 2, ..., 32$$

The survivor function and baseline hazard can be modelled using a Weibull distribution in order to make predictions of survival time for an individual and the median/mean survival time for different strata. Through computing mean squared errors and looking at a specific individual in our data set we find that our model could be used to predict a future patient's survival time.

In the Weibull regression model section, after combining Unknown and Regional for the stage, Other and Transverse for the subsite, and Coecum and Descending for the subsite, we found that individual survival time was highly depended on sex + stage_new + subsite_new + yydx + age. The existence of yydx was optional when considering the over-fitting problem.

$\mathrm{h}_i(t) = 6.13 \times 10^7 \times 0.7296t^{-1.7296}exp\{-0.1677 \times sex + 1.7142 \times stage[Distant] + 0.5906 \times stage[UnknownRegional] + 0.1083 \times subsite[OtherTransverse] + 0.0384 \times age - 0.0122 \times yydx\}$

$\mathrm{h}_i(t) = 0.0020 \times 0.7361t^{-1.7361}exp\{-0.1580 \times sex + 1.7172 \times stage[Distant] + 0.5876 \times stage[UnknownRegional] + 0.1059 \times subsite[OtherTransverse] + 0.0377 \times age\}$

Employing these two models to predict individual survival times in the test dataset, the model without yydx performed better than the model with yydx.

5 Discussion

The aim of the Cox regression model is to calculate the hazard ratio of different predictor variables. However, when it comes to estimating the survival time it's essential to fit a distribution to the survivor function and the baseline hazard. This however uses parametric techniques because the Cox regression model do not assume any particular distribution for the baseline hazard or the survivor function. Therefore, although this has the smallest mean square error and is the best fit to the data it might be best to also consider the Weibull model.

The Cox regression model after stratification contains 64 terms. Although this may be the best-fitting model for the data set, further statistical methods that haven't been considered in this report might be applied to simplify the interaction model to have fewer product terms to avoid over-fitting.

In the prediction section of the Weibull regression model, it was hard to decide which model was better fitted to the whole data set. Further action could include using the bootstrapping method. The fundamental concept behind bootstrapping involves creating multiple samples by selecting observations from the original data set repeatedly, allowing for replacement each time. We fit every sample to two models respectively and compare the MSE. If one model performs better among almost all samples, we can say it fits the original data better than the other. Another problem is that both of these two models do not fit when patients are quite young. It might be worth dividing the original data set into several groups according to age, and fitting them with different models.

Point predictions have been evaluated for the purposes of this report, but are not widely used in survival analysis because of the difficulty to make them accurately. Instead, survival models such as the Cox regression are typically used to identify factors that are associated with increased or decreased risk of an event occurring, rather than making individual predictions for survival times.

Another model we could have looked at is models which use machine learning such as the random forest model. This model combines the outputs of decision trees to reach a final result and builds a risk prediction model for survival analysis.

6 Reference

Collett, D. (2015). *Modelling survival data in medical research*, CRC press.

Clark, R. and B. M. Peck (2012). "*Examining the gender gap in life expectancy: a cross-national analysis, 1980–2005.*" Social Science Quarterly 93(3): 820-837.

Kleinbaum, D. G. and M. Klein (1996). *Survival analysis a self-learning text*, Springer.

Labianca, R., et al. (2010). "*Colon cancer.*" Critical reviews in oncology/hematology bootstrapping33.

O'Connell, J. B., et al. (2004). "*Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging.*" Journal of the National Cancer Institute 96(19): 1420-1425.

Zhang Z. (2016). *Parametric regression model for survival data: Weibull regression model as an example.* Annals of translational medicine, 4(24), 484.

7 Appendix

This section will show any extra plots which might be of interest.

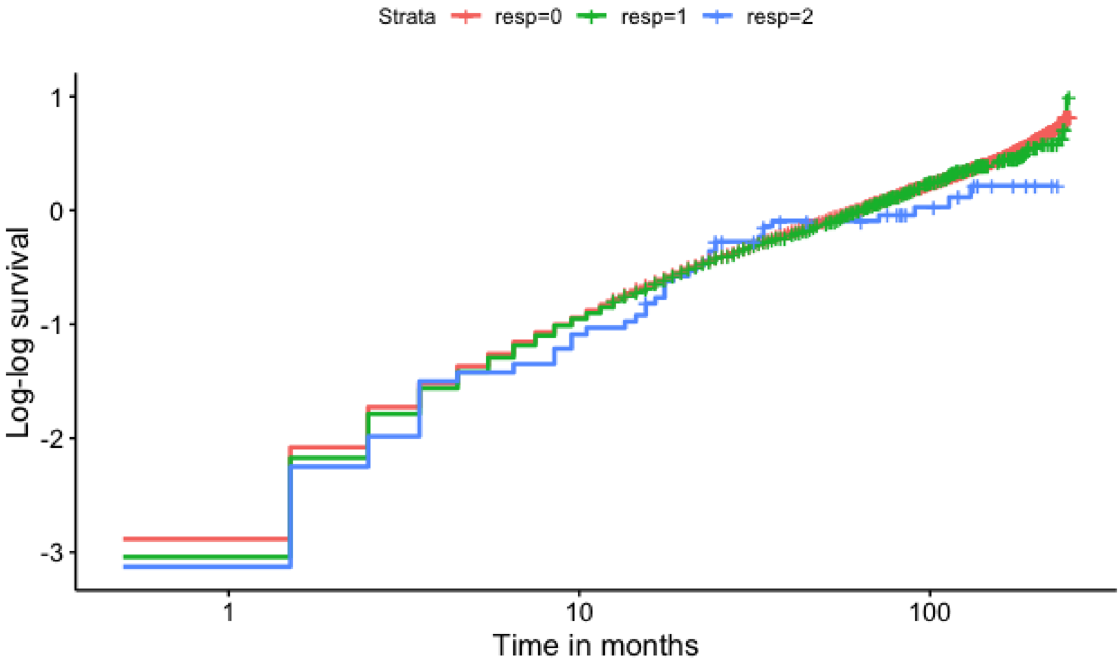


Figure 20: Log-log plot of respiration in Cox regression model

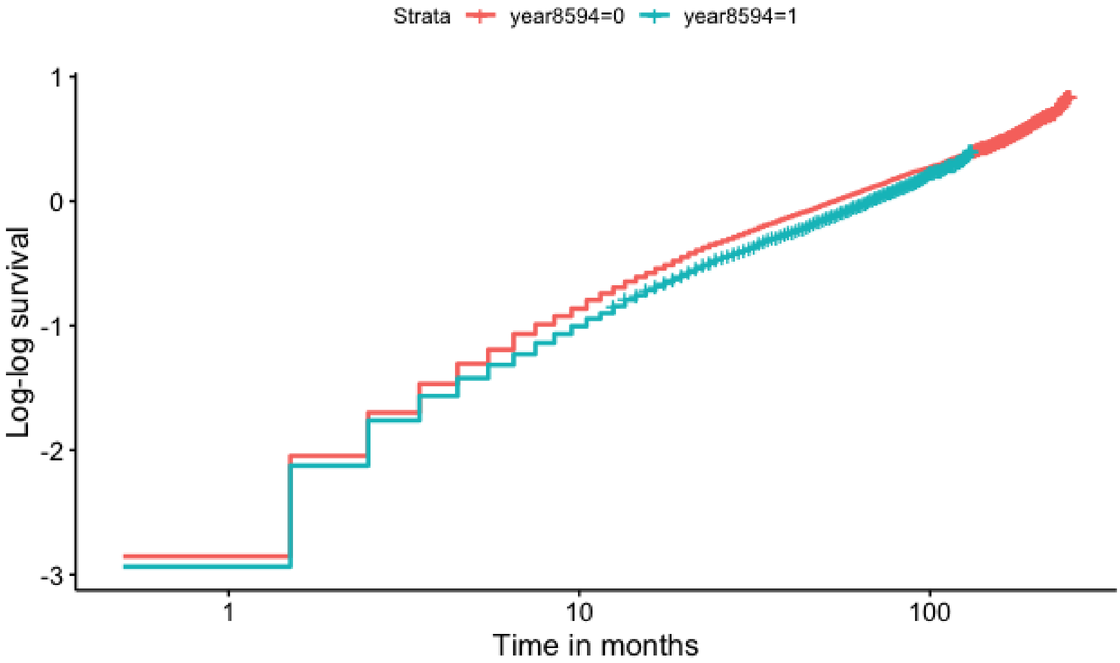


Figure 21: Log-log plot of year8594 in Cox regression model

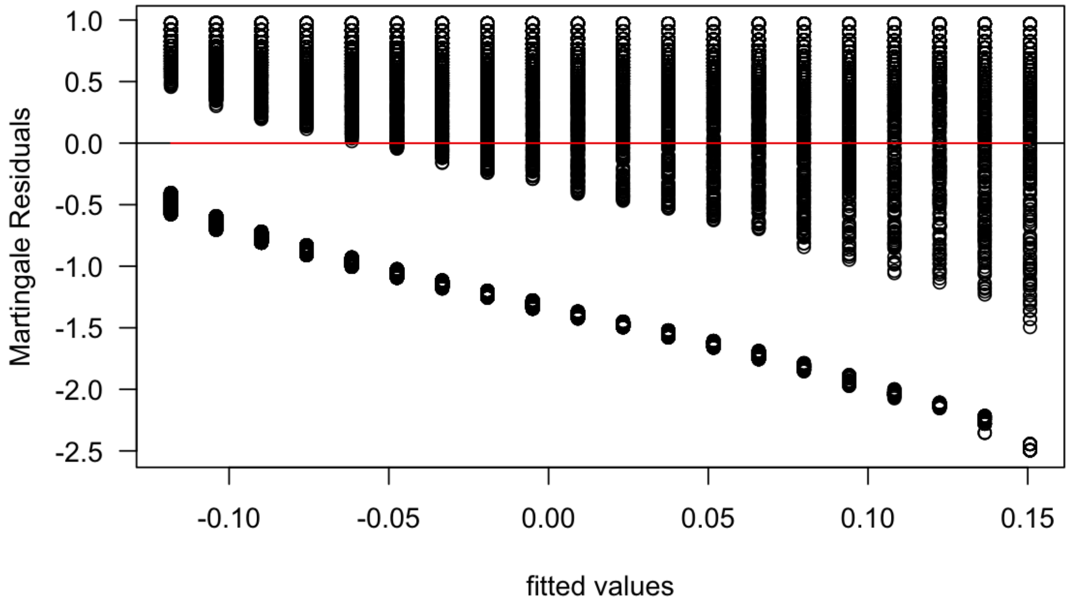


Figure 22: Martingale Residual plot of yydx

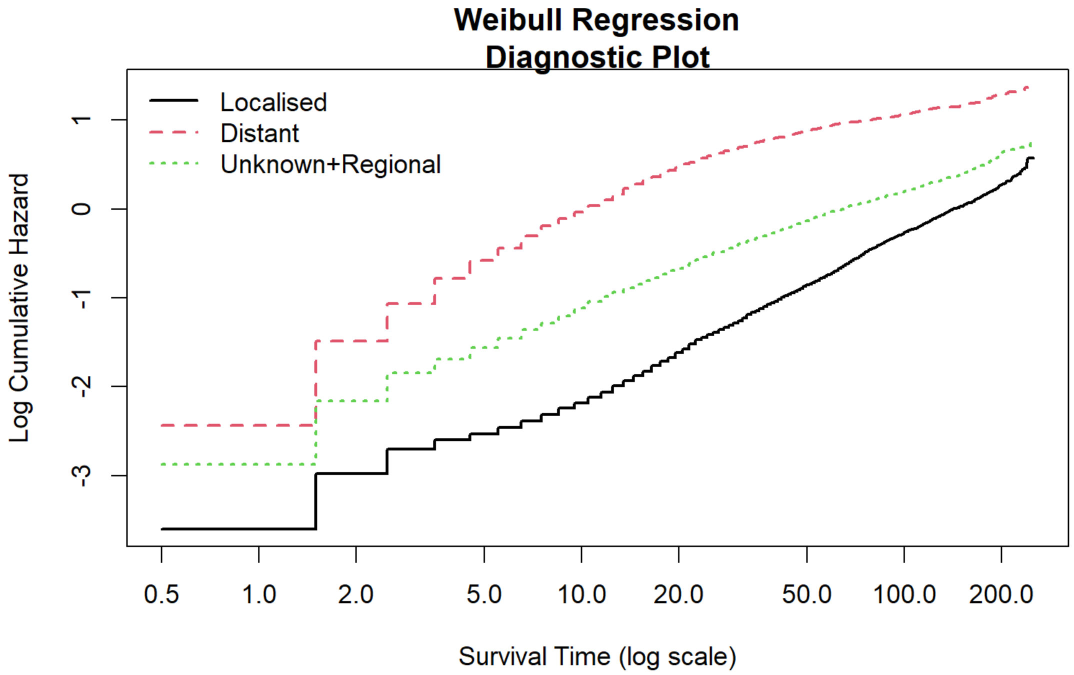


Figure 23: log-log plot of stage in Weibull model

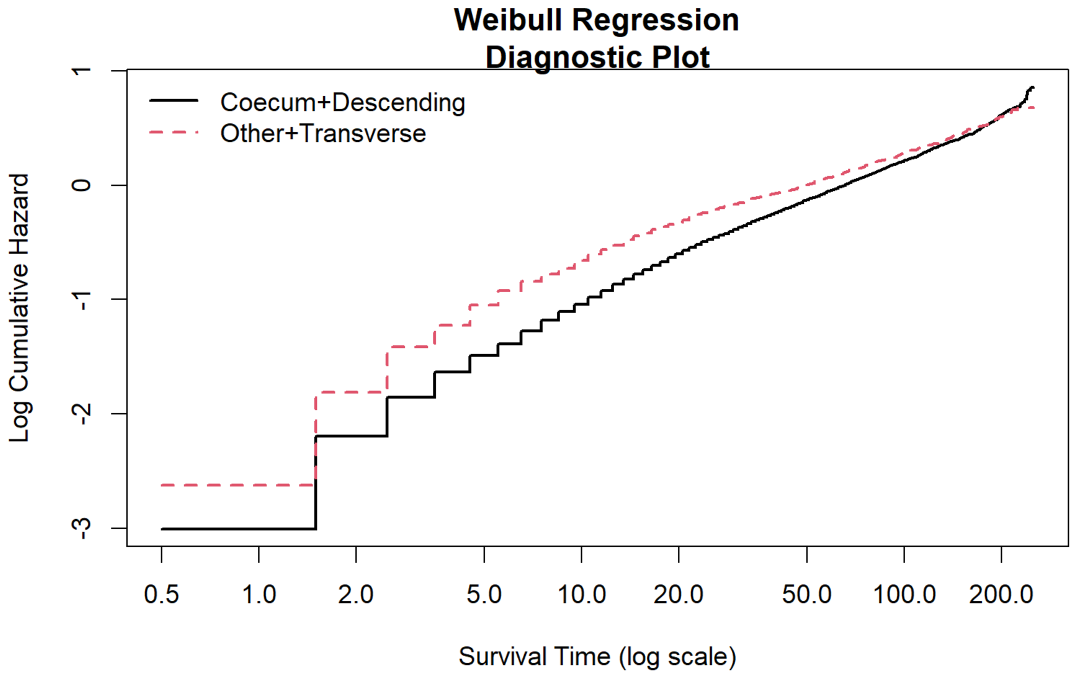


Figure 24: log-log plot of subsite in Weibull model

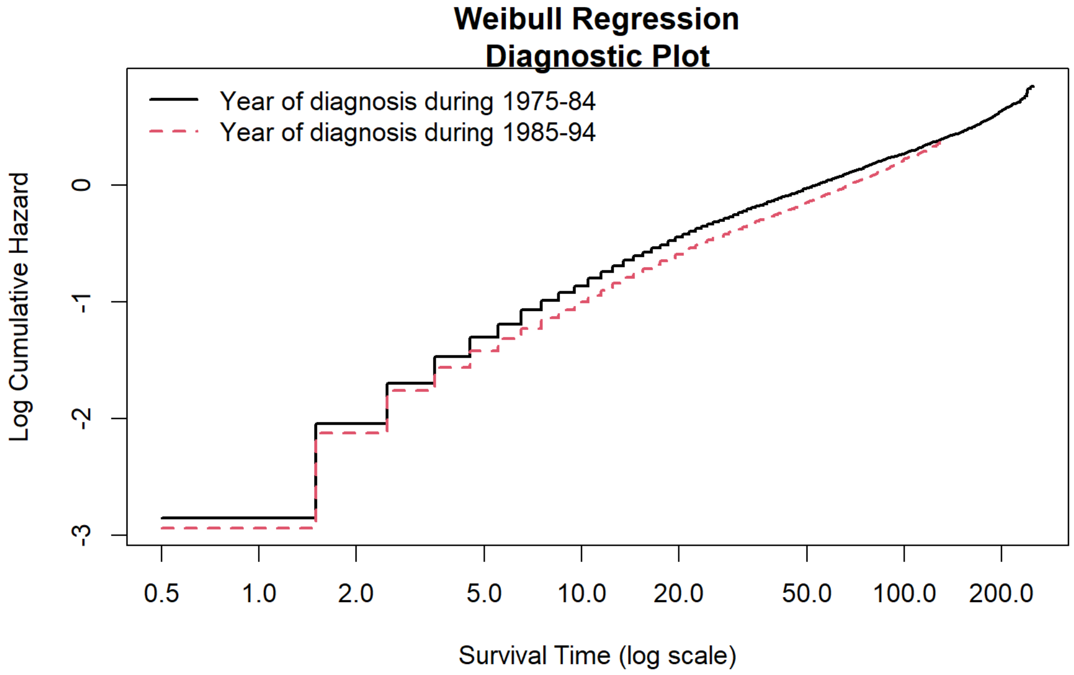


Figure 25: log-log plot of the year of diagnosis in Weibull model

Table 13: Summary result of Cox regression model

Parameters	Estimated coef
λ	7.788536e+11
γ	6.256439e-01
age	-5.615004e-02
age^2	6.902854e-04
yydx	-1.484541e-02