A statistical analysis of factors affecting the survival of patients with colon cancer

Alisdair Cole, Adam Corby, Aysha Dodhy, Jiaxuan Shi, Zhuotao Wang

Thursday 14th March, 2024

Abstract

In this report, we employed survival analysis to develop seven robust models for predicting the survival time of patients diagnosed with colon cancer, incorporating several prognostic variables. Assessing the suitability of each model using the standard measure of concordance, our findings reveal minimal discrepancies in accuracy across the models, suggesting that all models possess similar predictive capabilities. We found that the variables influencing a patient's survival time are their age, the tumour's anatomical subsite, their year of diagnosis and the clinical stage at diagnosis, with the cancer stage being most critical for patients' outcomes. By exploring the methodologies employed for each model, readers will gain comprehensive insights into our approach and findings.

Contents

1 Introduction

1.1 Restatement of the task

Colon cancer poses a significant challenge to public health globally, necessitating ongoing research to improve prognostic understanding and patient outcomes. In this project, we analyse data from 15,564 patients diagnosed with colon cancer in a Northern European country between 1975 and 1994. Our primary aim is to develop a statistical model to predict survival time based on various prognostic factors. We also seek to identify the most critical variables influencing survival prognosis and enable the prediction of individual patient outcomes. This report outlines our methodology, findings, and the implications of our model for clinical decision-making in colon cancer management when values of prognostic variables are known.

Before we begin devising a statistical model, we believe it would be important to gauge a general understanding of colon cancer and the general statistics of the disease. From this research we found some interesting statistics about prognosis based on sex, the tumour's anatomical subsite and the cancer's stage at diagnosis:

- The survival rate for males is approximately 3% lower than for females [4].
- The survival rate highly depends on the stage at which the cancer is diagnosed. The survival rate is considerably higher if the cancer is localised (approximately 91%) compared to if the cancer is distant (approximately 13%) [1].
- Ascending colon cancer is associated with the poorest survival outcome, whereas descending colon cancer is associated with the best survival outcome [7].

We expect to find that the data supports us in research. In this report, we will investigate which of these variables are the most important or most valuable when determining the prognosis. We will begin with an introduction to survival analysis and the random forest model, before exploring the Cox proportional hazards model. Then, we will explore some parametric models, before developing the random survival forest model. Each model will also have a section describing how we improved the model. Finally, we will compare each of these models against each other to determine which model is the best.

1.2 Background and literature review

1.2.1 Survival analysis

We begin with an introduction to survival analysis. Survival analysis is the analysis of the time until the occurrence of an event of interest. In this study, the 'event' will be a patient's death. Survival analysis involves modeling the probability distribution of the time until an event occurs, which is referred to as the survival time. This distribution is typically represented by the survival function, which gives the probability that an individual survives beyond a certain time point. The survival function is complementary to the cumulative

distribution function and is often denoted as $S(t)$, where t represents time. There are three key concept in survival analysis:

- **The Survival Function,** $S(t)$: This is a function which gives the probability that an individual survives beyond a certain time, t. The survival function is non-increasing, meaning that as time progresses, the probability of survival decreases or stays the same.
- **The Hazard Function,** $\lambda(t)$: This represents the instantaneous rate at which events occur at time t, given survival up to that time. It describes the risk of experiencing the event of interest at any given moment in time. This can vary over time, reflecting changes in the risk of experiencing the event as time progresses. The higher values of the hazard function indicate a higher instantaneous risk of the event occurring.
- **Survival Models**: Various statistical models are used to analyse survival data, including the Kaplan-Meier estimator, Cox proportional hazards models, parametric survival models, and more. These models allow researchers to estimate survival probabilities; identify risk factors associated with the event of interest; and make predictions about future outcomes.

One of the key reasons that we choose survival analysis to approach this analysis is due to how it handles censored data. Censored data is data which we don't have full information for, and, in the case of this study, takes the form of right-censored data. This is data which have been cut short by the end of the study, before the measured event has occurred. In this report, this refers to patients who are alive at the end of the study, so we cannot know when they die. See Figure 1 for a visual representation of right censored data. Patient 1 dies before the end of the study, so their data are complete. Patient 2, however, dies occurs after the study has ended, so the data on Patient 2 data are right-censored.

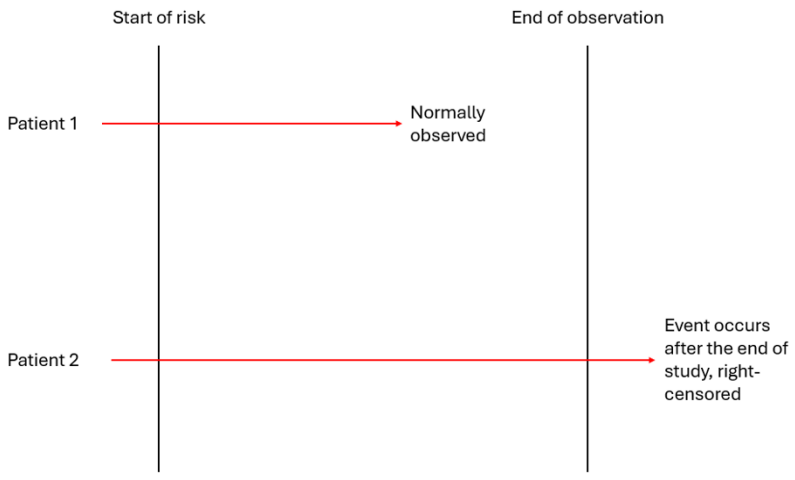


Figure 1: A visual representation of right-censored data.

1.2.2 Machine learning and the random forest model

Classification And Regression Tree (CART) models form a class of statistical models that operate by recursively splitting a dataset into subsets based on the most significant predictors. This process forms a large decision tree, with each node representing a prediction outcome. Through this approach, CART models allow the identification of key variables and the interaction between different predictors, making them a key tool in the field of machine learning.

Figure 2 illustrates how a CART model works by plotting a decision tree. In this example, the first decision is made, leading to the formation of two sub training sets. Then, the decision is made again which leads to four sub-sub training sets, and this process continues in the same way until the optimal decision is found.

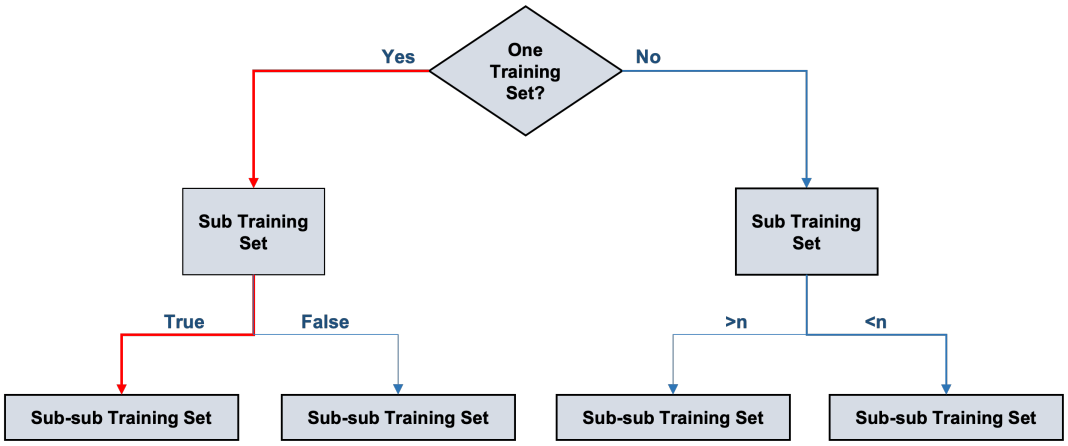


Figure 2: An example of a decision tree from a CART model.

The random forest (RF) model is derived from the CART class of models. It is a machine-learning algorithm that combines the output of multiple decision trees to reach a single output. The random survival forest (RSF) model is an adaptation of this technique, which we can use to aid our analysis since it is specially designed to address right-censored data. In the framework of survival analysis, the RSF model provides a number of additional benefits and features. It is non-parametric, meaning it does not require assumptions about the shape of the survival curve or the proportional hazards assumption, offering flexibility in modeling complex survival data [11].

The RSF model operates by constructing a number of decision trees, with censoring taken into account - see Figure 3. At each node of these trees, a subset of predictors is randomly selected as candidates for splitting. The process then aggregates across the survival predictions of the entire forest of trees, to provide a comprehensive estimate of survival probabilities. Hence, the random survival forest model uses the collective insights of multiple decision trees to enhance the model's ability to uncover non-linear relationships and interactions and ensure a robust analysis of the survival data.

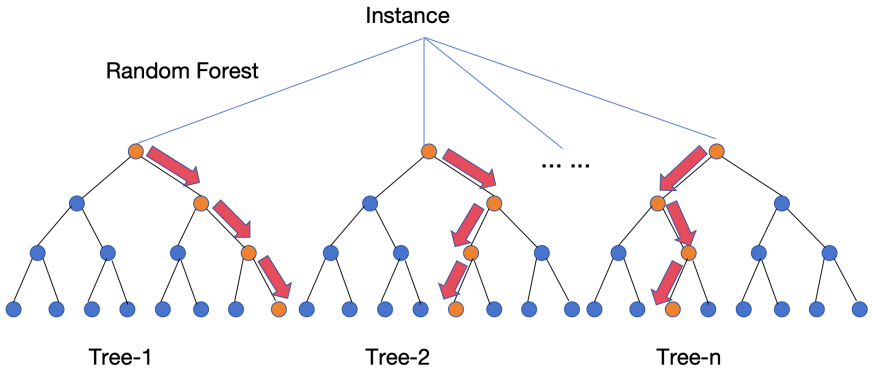


Figure 3: An example of a subset of decision trees created from a random forest model.

1.3 Data processing

1.3.1 Assumptions

To simplify the problem, we first made the following three basic assumptions about the data:

1. The observations were sampled at random from the population.
2. All records of observations were accurate.
3. The significant medical development of the TNM staging system (standard for classifying the anatomical extent of the spread of cancer) in 1984 was assumed to have affected the data.

1.3.2 Data structure

Name	Type	Variable Content
sex	character	Female/Male
age	integer	Age at diagnosis, e.g., 77, 78, 78
stage	character	Unknown/Localised/Regional/Distant
mmdx	integer	Month of diagnosis, e.g., 3 = March, 7 = August, etc.
yydx	integer	Year of diagnosis, e.g., 1977, 1978, 1978
surv_mm	integer	Survival time in months, e.g., 16.5, 82.5, 1.5
surv_yy	integer	Survival time in years, e.g., 1.5, 6.5, 0.5
status	character	Alive/Died of cancer/Died with cancer/Lost to follow-up
subsite	character	Transverse/Coecum and ascending/Descending and sigmoid/Other and NOS
year8594	character	Diagnosed 75-84/Diagnosed 85-94 (can take two possible values)
id	integer	Identification Number, e.g., 1/2/3/4/5
resp	integer	0/1 (0 if patient does not have a respiratory illness, 1 if they do)

Table 1: An overview of the variables in the dataset.

Table 1 characterises our data, showing variable names, data types and content variations.

1.3.3 Data handling

We have identified instances where certain variables overlap in their representation within our dataset. For example, both 'surv_mm' and 'surv_yy' indicate survival length, essentially conveying the same information. In our analysis, we've opted to utilize 'surv_mm' for its potential to offer a more precise measurement. Similarly, 'year8594', 'yydx', and 'mmdx' all represent the diagnosis time. To maintain coherence in our analysis, we've selected 'year8594' as the primary diagnosis date to explore potential relationships between the year period and survival time. Furthermore, we have excluded the 'id' variable from our models as it primarily serves as a serial number for each patient.

Consequently, our starting covariates will consist of 'sex', 'age', 'stage', 'subsite', 'year8594', and 'resp', while our response variables will encompass 'status' and 'surv_mm'.

Additionally, we have made the decision to remove four data points labelled as 'Lost to follow up'. Given the absence of a clear resolution method for this issue within the dataset and the minimal impact of these data points, we deemed it sensible to eliminate them to uphold data integrity and enhance the accuracy of our analysis.

2 Methods and results

This section explains the non-parametric, semi-parametric and parametric models we have used in this report in detail, and their corresponding results.

A key metric we used to evaluate the performance of each model will be the concordance level. This is a commonly used metric in survival analysis, since it encompasses both observed events and censored cases. The Concordance Index (C-index) quantifies the model's ability to correctly rank pairs of subjects based on their predicted survival times. Specifically, for any randomly chosen pair of individuals from our study population, the C-index reflects the probability that the patient who dies first also has the higher predicted risk of the event as estimated by the model. Hence, a concordance of 0.5 suggests that the model's predictions are no better than random chance in ranking survival times, indicating no predictive discrimination; and a concordance of 1 denotes perfect predictive accuracy. Values between 0.5 and 1 therefore indicate varying degrees of predictive accuracy, with higher values representing better model performance.

2.1 The Kaplan-Meier estimator

Using the definition of the survival function outlined in Section 1.2.1 above, we defined our survival function, $S(t)$, as a function that gives the probability that a patient will survive past a certain time, t, in months.

We used the Kaplan-Meier product-limit (PL) estimator to obtain an estimate for $S(t)$. This is a non-parametric statistic, particularly useful for analysing time-to-event data, as we had in this instance with colon cancer patient deaths. The use of these non-parametric methods is generally the starting point of most contemporary analyses [9].

Defining our survival event as the death of a patient, this estimator is given by:

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

where t_i is a distinct time when at least one death occurred, d_i is the number of deaths that happened at time t_i and n_i is the number of patients who have survived at time t_i.

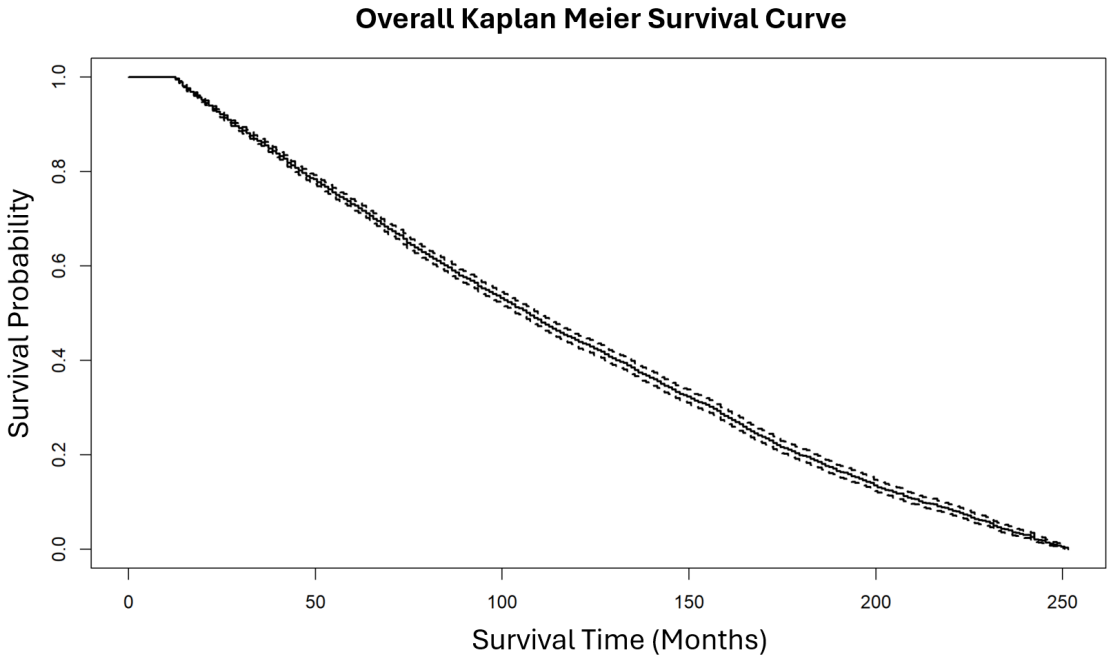
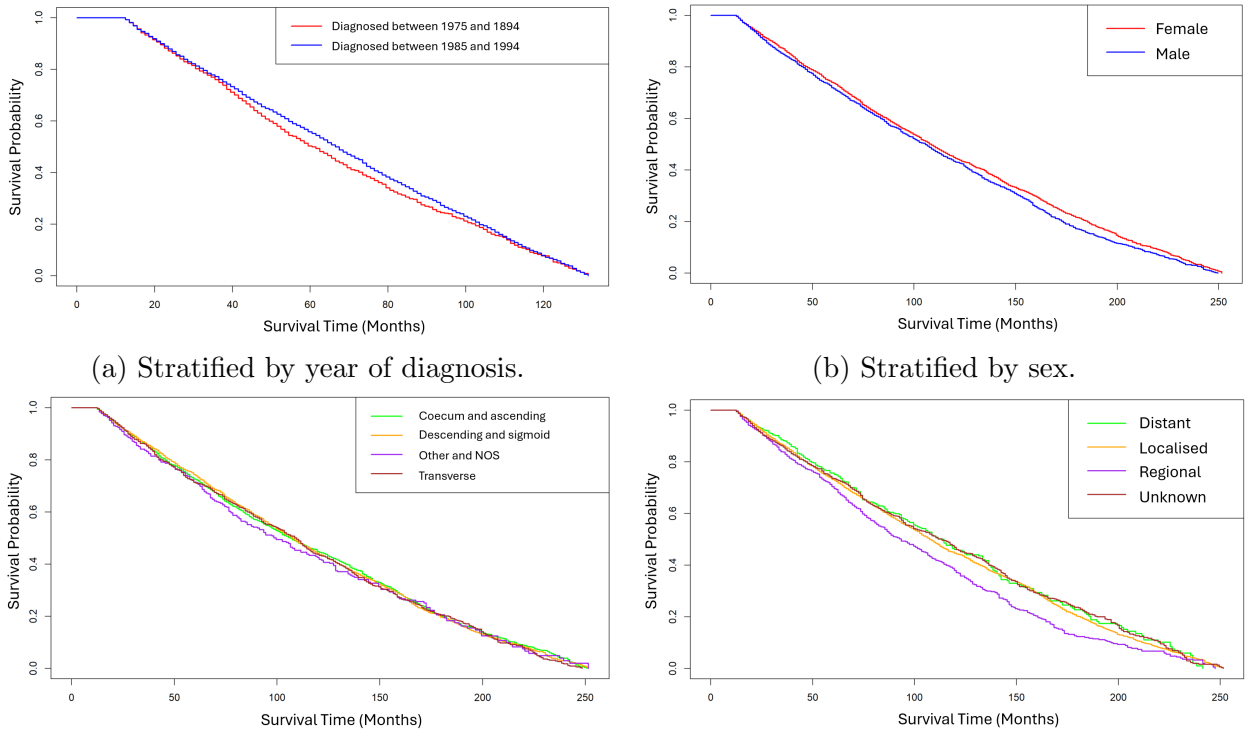


Figure 4: A plot showing the overall Kaplan-Meier survival curve for 15,564 colon cancer patients.

Figure 4 shows the overall Kaplan-Meier survival curve for the data, with point-wise 95% confidence intervals also plotted. Stratified Kaplan-Meier curves can help us further explore the data - that is, creating separate Kaplan-Meier curves for different subgroups based on prognostic variables.



(a) Stratified by year of diagnosis.

(b) Stratified by sex.

(c) Stratified by subsite.

(d) Stratified by stage.

Figure 5: Plots showing stratified Kaplan-Meier survival curves for 15,564 colon cancer patients.

Figure 5a shows two distinct Kaplan-Meier survival curves, based on an indicator variable for the year of diagnosis. That is, the red survival curve contains patients who were diagnosed between 1975 and 1984, and the blue survival curve contains data from patients who were diagnosed between 1985 and 1994. By inspection, we noticed that the difference between the curves increases up to a point of around 5 years post diagnosis, before the survival curves begin to converge.

The next step was to ascertain whether the two survival curves differ in a way that is statistically significant. In other words, is there a reason to believe that a patient diagnosed between 1985 and 1994 has a far higher chance of surviving than a patient diagnosed ten years earlier? A log-rank test can be used for this, with the null hypothesis being that the survival function for the two groups is the same. A p-value of $< 2e - 16$ was obtained as a result, indicating strong evidence to reject the null hypothesis. This suggested that there was a significant variation in survival outcomes that is related to the year of diagnosis, and, more precisely, the difference between the two time periods.

Similarly, Figure 5b shows the survival probability plotted against survival months for male and female patients. Although there is very little difference between the survival probability early on in the patients diagnosis, after approximately 100 months there is a larger difference between survival probability of a male versus a female. This figure shows that the survival rate of a male patient is lower than the survival rate of a female, which supports our initial

research. To test whether the difference in these curves is significant, we used a log-rank test, as above. Again taking the null hypothesis as the survival function being identical for the two groups, this test gave a p-value of $2e-05$. This low p-value meant there was strong evidence to reject H_0 and hence suggested that there is a significant difference in survival outcomes between male and female patients.

Figure 5c shows four Kaplan-Meier survival curves, based on the anatomical subsite of the tumour. All of these curves looked fairly similar, and this observation was backed up by performing another log-rank test which produced a p-value of 0.8. This indicated that the anatomical subsite alone may not be a strong predictor of survival for the patients in the study, and hence may have a limited role in explaining survival outcomes in the models developed in Sections 2.2and2.3.

Finally, Figure 5d shows the survival probability based on the stage of cancer plotted against the survival time. We noticed that the survival probability for regional cancer is lower than all other stages of cancer. To begin with, all stages follow a similar trend until after approximately 50 months after diagnosis, where the trend for regional stage cancer defers and the survival probability declines more rapidly than the other stages. This contradicted our initial research where we would expect for distant stage cancer to have a lower survival probability. It was also clear from Figure 6 that distant, localised and unknown stage cancer have very similar survival probabilities. A log-rank test on this stratified plot, with analogous hypotheses to above, gave a p-value of $5e-05$, which indicated strong statistical significance that the stage of cancer is an important variable to use to explain survival outcomes.

The above Kaplan-Meier curves give a good indication of the effect the variables have on survival probability. These curves have provided us with useful information on potentially the most important variables that effect a patient's prognosis. It is clear that the different subsites had very little difference in the survival probability, indicating that this may be a less important factor when considering a patient's diagnosis. Conversely, the KM curves for 'sex', 'stage' and 'year8594' showed that some variables behave differently and have a lower survival probability than others. We will take these findings forward when devising our models.

It is also important to note that it was clear from the KM curve stratified by year of diagnosis, that there were significant disparities in survival between patients diagnosed in different time periods. This suggested that the influence of evolving medical practices and treatment modalities over time. Medicine has continued to evolve since this study and therefore this may have an influence on the accuracy of our models if they were to be used in the present day.

2.2 The Cox proportional hazards regression model

We attempted to fit a Cox proportional hazards regression model to our data, as we wanted to assess the effect of several risk factors on survival time simultaneously. Firstly we tried the Cox proportional hazards model with fixed covariates. Based on these results, which will

be detailed in the following Section, we formed a similar Cox model, but with time-varying covariates.

Before building the models, we first pre-processed the data. As detailed in Section 1.3.2, we would include the following six varibales as covariates: sex(1=Male, 2=Female), age, stage(0=Unknown, 1=Localised, 2=Regional, 3=Distant), subsite(0=Descending and sigmoid, 1=Coecum and ascending, 2=Other and NOS, 3=Transverse), year8594(0=Diagnosed 75-84, 1=Diagnosed 85-94) and resp(0=none,1=respiratory complications at entry). Note that when fitting these models, the covariates must be numerically defined, hence the departure from the notation used in Table 1. Also, 98 observations have missing values for the variable resp, which is less than 0.7% of the entire dataset for our study. As a result, we excluded these patients from this point onwards.

2.2.1 The Cox proportional hazards model with fixed covariates

The basic Cox model with fixed covariates is expressed as:

$$h_i(t) = h_0(t)exp(\beta_1 x_{i1} + ... + \beta_k x_{ik}) \tag{1}$$

where t represents the survival time; $h(t)$ is determined by a set of k covariates $(x_1, x_2, ..., x_k)$; $h_0(t)$ is the baseline hazard function and can be interpreted as the hazard function for a subject whose covariates all have the value of zero, and the coefficients β_i measure the impact of each covariate on the response variables:

- $\beta_i > 0$: the covariate is associated with an increased hazard of the patient dying;
- $\beta_i < 0$: the covariate is associated with a decreased hazard of the patient dying;
- $\beta_i = 0$: no association between covariate and patients' survival chances.

After building this initial model, we attempted to identify the most important variables in terms of survival chances. This could be done by comparing the coefficients β_i and predicting the survival for a given individual when the value of prognostic variables in this expression is known.
We first estimated the time to death with fixed covariates and modeled the survival time in months between diagnosis and death of cancer or censoring by the end of the study (i.e. the last date of contact).

We fitted the model with all six covariates and obtained the results shown in Table 2. In this table, 'coef' shows the estimate of the coefficient β_i for each variable, followed by 'exp(coef)', the exponential of the coefficient, which represents the multiplicative effects on the hazard. This is often referred to as the hazard ratio. The hazard ratio was found to be less than 1 for the 'sex', 'year8594', and 'resp' variables. Therefore, these results suggested that female patients; patients diagnosed between 1985 and 1994; and patients with respiratory complications at entry all had a lower hazard of dying. Similarly, we saw that the hazard ratios for age, subsite, and stage variables were all larger than 1, indicating these variables are associated with worse patient survival outcomes.

	coef	exp(coef)	z	p
sex	-0.04	0.96	-1.63	0.10
age	0.03	1.03	22.32	< 0.005
subsite	0.06	1.06	5.33	< 0.005
year8594	-0.21	0.81	-8.26	< 0.005
stage	0.72	2.05	53.94	< 0.005
resp	-0.04	0.96	-0.92	0.36

Table 2: A summary of the basic Cox model with fixed covariates.

The Akaike Information Criterion (AIC) tests how well a model fits a dataset, without over-fitting it. The AIC score rewards models that achieve a high goodness-of-fit score and penalizes them if they become overly complex. The model with the lower AIC score is expected to strike a superior balance between its ability to fit the data set and its ability to avoid over-fitting the data set [3]. The AIC statistic is calculated as:

$$AIC = -2(logL) + 2(c + p + 1)$$

where $-2logL$ is twice the negative of the log-likelihood; c is the number of covariates in the model; and p is the number of structural parameters in the model. The results are summarised in Table 3.

Number of Variables	Best AIC Score	Best Selected Candidates
1	148766.7058	stage
2	148087.2646	stage, age
3	148009.9602	stage, age, year8594
4	147970.9568	stage, age, year8594, subsite
5	147969.6221	stage, age, year8594, subsite, sex
6	147969.9388	stage, age, year8594, subsite, sex, resp

Table 3: A table comparing AIC scores across a selection of models, each with an increasing number of variables.

The smallest AIC score was obtained when selecting five variables: 'stage', 'age', 'year8594', 'subsite', and 'sex'. Fitting the model with these five variables again produced the following results:

	coef	exp(coef)	z	p
sex	-0.04	0.96	-1.61	0.11
age	0.03	1.03	22.33	< 0.005
subsite	0.06	1.06	5.33	< 0.005
year8594	-0.21	0.81	-8.25	< 0.005
stage	0.72	2.05	53.93	< 0.005

Table 4: A summary of the best basic Cox model with fixed covariates.

Figure 6 shows the log of the hazard ratio, i.e. β, for each of the five covariates in the model.

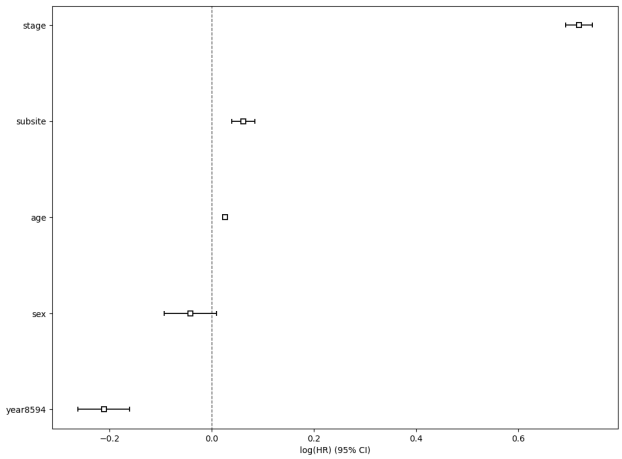
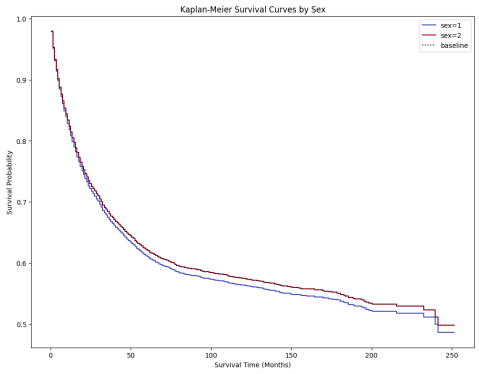
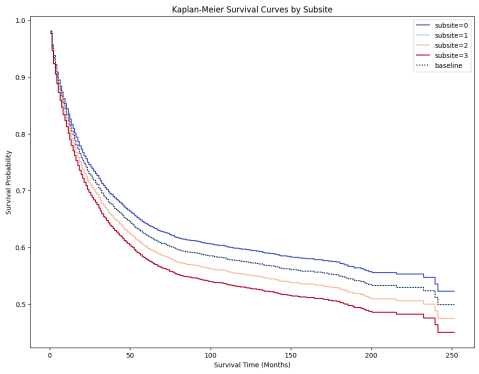


Figure 6: A plot showing the log of the hazard ratio for each of the five covariates in the best basic Cox model with fixed covariates

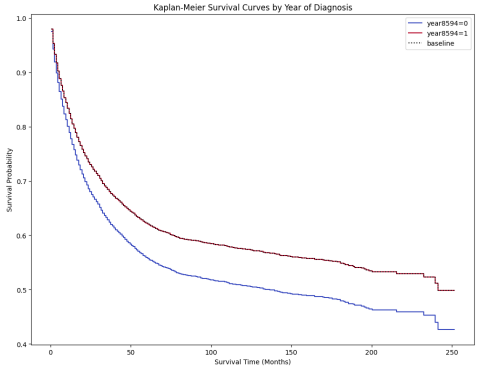
The revised Kaplan-Meier plots in Figure 7 below show the partial effects of different covariates on survival outcomes.



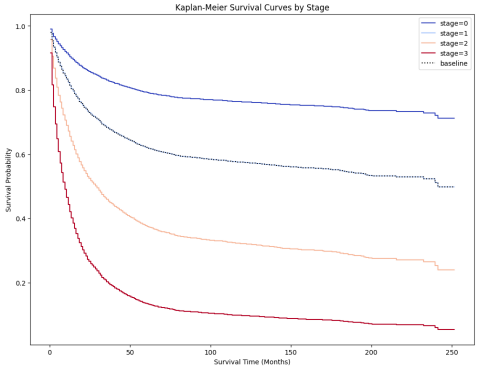
(a) Stratified by sex.



(b) Stratified by subsite.



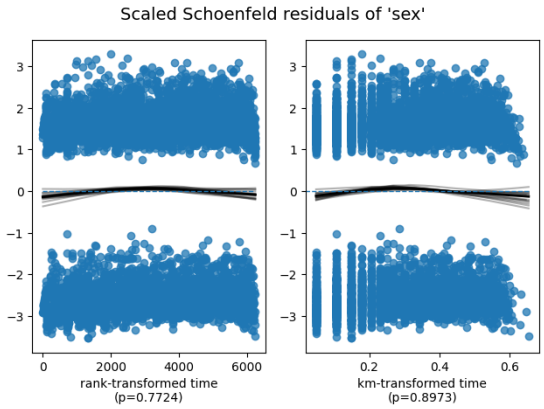
(c) Stratified by year of diagnosis.



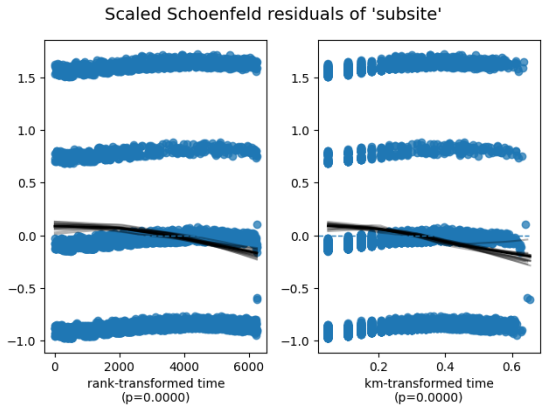
(d) Stratified by stage.

Figure 7: Revised Kaplan-Meier survival curves based on the Cox model with fixed covariates.

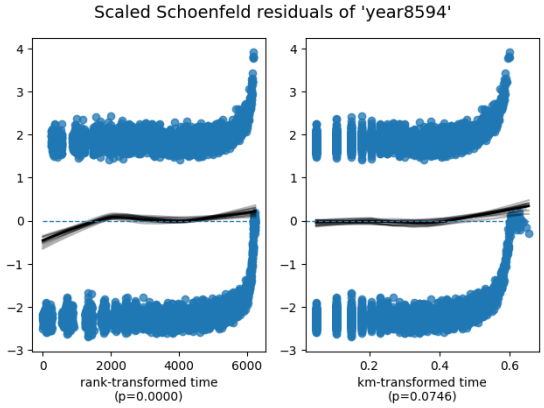
The Cox proportional hazards model assumes that the hazard ratio for any two predictors is constant over time. In order to validate this assumption, we needed to observe the Schoenfeld residuals. The plots of these Schoenfeld residuals against time would reveal whether a particular coefficient from a covariate was time-dependent. If all covariates are time-independent, then the assumptions would have been satisfied.



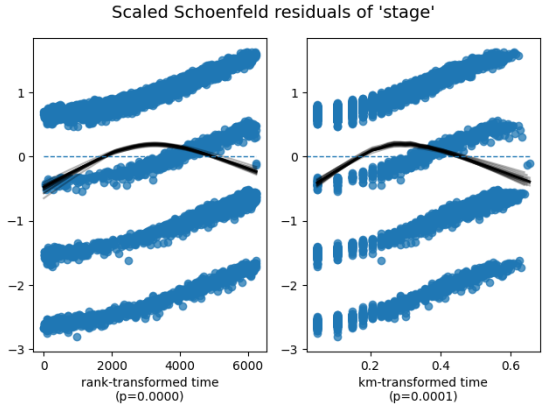
(a) Schoenfeld residual plot of sex.



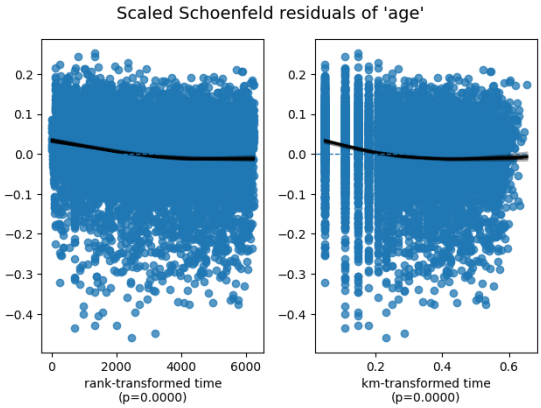
(b) Schoenfeld residual plot of subsite.



(c) Schoenfeld residual plot of year of diagnosis.



(d) Schoenfeld residual plot of stage.

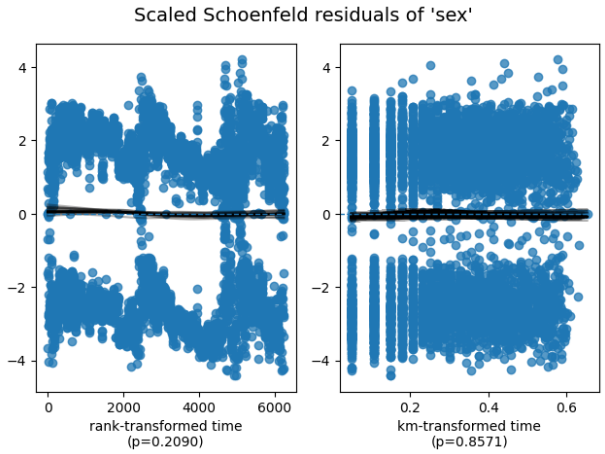


(e) Schoenfeld residual plot of age.

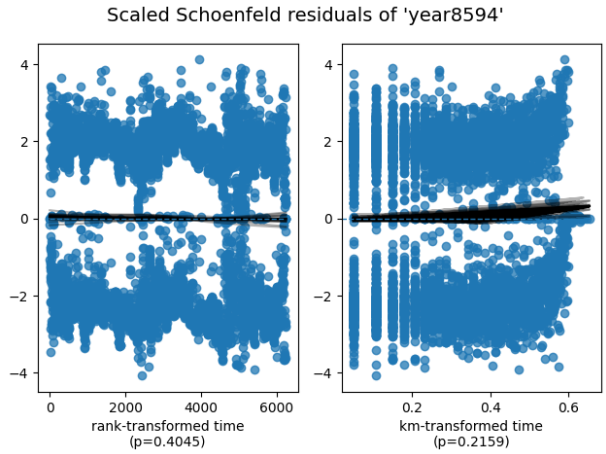
Figure 8: Schoenfeld residual plots of the best basic cox model with fixed covariates.

From these plots, we found the variables 'stage', 'subsite', 'age', and 'year8594' had non-proportional hazards. That means these four covariates were found to be time-dependent, violating a key assumption of the model. To deal with this, we divided the data into strata. That means after stratification, each stratum would be allowed to have a different baseline hazard function. Along with this, the coefficients could be assumed constant across the different strata, so we did not need to satisfy the proportional hazards assumption for the stratified covariates.

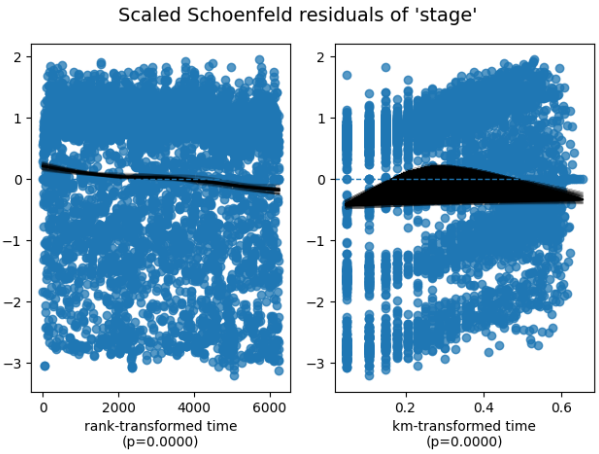
We first stratified the data based on the 'age' and 'subsite' variables, calling it Stratified Model 1. We used Schoenfeld residuals again to check the assumptions. Since we treat variable 'age' and 'subsite' as time-dependent, we can only plot the Schoenfeld residuals plot of 'sex','year8594' and 'stage'. However, the stage covariate still violated the proportional hazard assumption, as shown in Figure 9c.



(a) Schoenfeld residual plot of sex (Stratified Model 1).



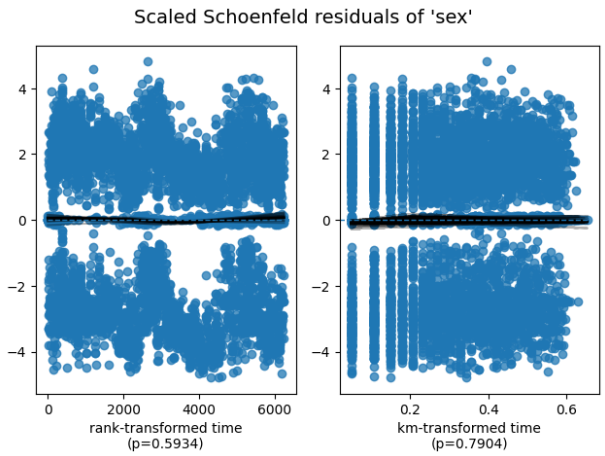
(b) Schoenfeld residual plot of year of diagnosis (Stratified Model 1).



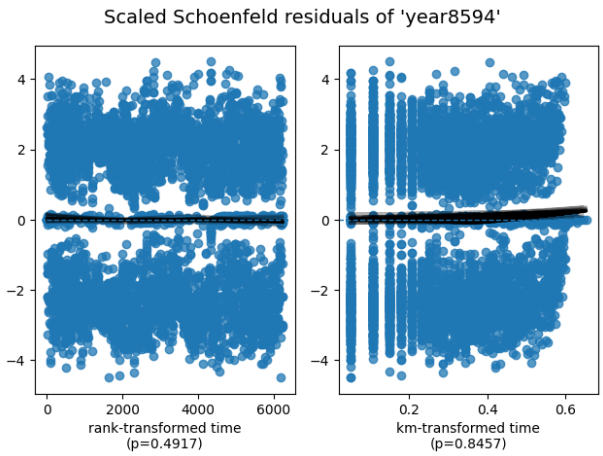
(c) Schoenfeld residual plot of stage (Stratified Model 1).

Figure 9: Schoenfeld residual plots of the Stratified Cox Model 1 with fixed covariates.

We then stratified data based on the 'age', 'subsite' and 'stage' variables, calling it Stratified Model 2 and once again used Schoenfeld residuals plot of the other two variables, 'sex' and 'year8594' to check the assumptions. However, the concordance for this model is not so high that we decided to use another method called time-varying covariates.



(a) Schoenfeld residual plot of sex (Stratified Model 2).



(b) Schoenfeld residual plot of year of diagnosis (Stratified Model 2).

Figure 10: Schoenfeld residual plots of the Stratified Cox Model 2 with fixed covariates.

Thus, in this section, we tried to satisfy the assumption of Cox-proportional hazards model by stratification, however, when fitting the model, the concordance result was not so good, thus, we need to develop our method in the next section using time-varying covariates.

2.2.2 The Cox proportional hazards model with time-varying covariates

From the assumption tests of the Cox proportional hazards model with fixed covariates, we found that the effects of some variables change over time. So, the next logical step was for us to fit a model using time-varying covariates.

Estimating a Cox proportional hazards model with time-varying covariates takes the following form:

$$h_i(t) = h_0(t)exp\{\beta_1 x_{i1} + \beta_2 x_{i2}(t) + ... + \beta_k x_{ik}\} \tag{2}$$

The model states that the hazard at time t depends on the value of time-fixed covariates (e.g., x_1), and on the value of the time-varying covariates (e.g., x_2) at time t.

We used the five covariates from the best basic Cox model with fixed covariates, and the results are shown in Table 5.

	coef	exp(coef)	z	p
sex	-0.04	0.96	-1.80	0.07
age	0.03	1.03	25.38	< 0.005
subsite	0.07	1.07	6.52	< 0.005
year8594	-0.20	0.82	-8.96	< 0.005
stage	0.80	2.23	53.56	< 0.005

Table 5: A summary of the Cox model with time-varying covariates.

For this model, the log of the hazard ratio (i.e., β) was less than 0 for the covariates 'sex' and 'year8594'. Therefore, patients who are female and patients who were diagnosed between 1985 and 1994 had a lower hazard of dying. Similarly, the coefficient values for 'age', 'subsite' and 'stage' variables were all larger than 0, indicating these variables are associated with worse patient survival outcomes.

This model identified the most important variables by evaluating the p-value and we found the most important variables in terms of prognosis for survival to be 'age','subsite', 'stage', and 'year8594'.

2.3 Parametric models

2.3.1 The Weibull model

As an alternative to the semi-parametric Cox model, we also used the parametric Weibull model. The Weibull model is recommended as a choice of parametric models for situations like the data we had - studies on the survival of patients and patient deaths[9].

The key difference between the Weibull model and the Cox model lies in the flexibility of the hazard function. The Cox model does not assume a specific functional form for the hazard, allowing it to adapt to different survival patterns. By contrast, the Weibull model assumes that the baseline hazard is characterized as monotonic. The hazard for the Weibull distribution is specified as follows:

$$h(t) = p\lambda t^{(p-1)} \qquad t \geq 0,\ p > 0, \text{ and } \lambda > 0$$

We checked whether the Weibull model was suitable by plotting the logarithm of the negative logarithm of the survival function against the logarithm of the time, and if this plot was roughly linear, we could use the Weibull distribution [5].

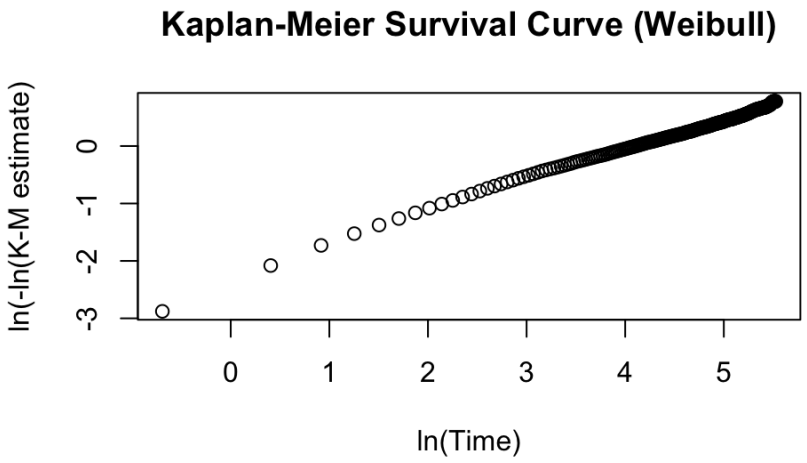


Figure 11: A plot of the log–negative log survival against the log of failure time.

The plot in Figure 11, which uses the Kaplan-Meier estimate, follows a linear trend, indicating the Weibull model was appropriate for our data. Hence, we confirmed that we could use the Weibull parameterisation, and could begin to develop this model further.

2.3.2 Developing the Weibull model

Our initial model using this parametric approach included every variable provided in our dataset. This helped ensure no variables with significant explanatory power were overlooked. The model is summarised in Table 6 below. Note that not all variables are featured in this table, as some variables were treated as a 0 for the purposes of modelling. For example, sex (Female) was treated as a 0, and the values for Sex (Male) were fitted based on that.

Variable	**Coefficient**	**P-value**
Intercept	7.464	≈ 0
Sex (Male)	-0.217	≈ 0
Age	-0.05	≈ 0
Stage (Localised)	0.859	≈ 0
Stage (Regional)	0.097	0.067
Stage (Distant)	-1.48	≈ 0
Subsite (Coecum and ascending)	0.138	≈ 0
Subsite (Descending and sigmoid)	0.189	≈ 0
Subsite (Other and NOS)	0.077	0.204
Diagnosed 1985-1994	0.121	≈ 0
Respiratory issues	0.059	0.18

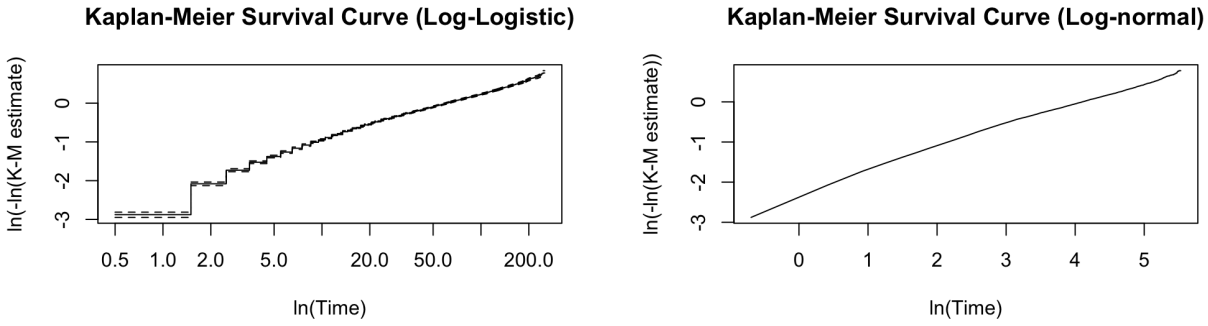
Table 6: A summary of the variables in the Weibull model.

This model had a log-likelihood of -49976.5, a concordance of 0.7284, and an AIC score

of 99977.09, both of which were used as baseline scores when considering potential improvements. Attempting to improve the model, we first removed the resp variable, due to its high p-value. However, this actually reduced the log-likelihood to -50278.6, increased the AIC score to 100579.3, and the concordance remained on 0.7284. This reduction in performance was mirrored if we remove other variables, so we could not improve this model further, and the model detailed in Table 6 is the best fitting Weibull model for the data.

2.3.3 The log-normal and log-logistic model

Additionally, we attempted to fit log-logistic and log-normal models. These are both parametric models that allow for non-monotonic hazard functions, unlike the Weibull model. Since the log-normal and log-logistic distributions are similar, they produced very similar results.



(a) A plot of the log–log survival against the log of the failure time.

(b) A plot of the log–log survival against the log of the failure time.

Figure 12: Kaplan-Meier survival curves for checking the suitability of the log-logistic and log-normal models.

Figure 12 contains two subplots of Kaplan-Meier curves, used to assess the suitability of each of these models. The plots in both Figure 12a and Figure 12b follow a linear trend, which indicated that both models were appropriate for this data and thus allowed us to proceed with attempting to fit these models.

2.3.4 Developing the log-normal model

In a similar way to how the Weibull model was fitted in Section 2.3.1, we fitted a log-normal model consisting of all possible variables. The model is summarised in Table 7 below. Again, for the same reason as above, not all variables are featured in this table.

Variable	Coefficient	P-value
Sex (Male)	–0.20053	≈ 0
Age	-0.05006	≈ 0
Stage (Localised)	1.14093	≈ 0
Stage (Regional)	0.36633	≈ 0
Stage (Distant)	-1.33502	≈ 0
Subsite (Coecum and ascending)	0.21454	≈ 0
Subsite (Descending and sigmoid)	0.24523	≈ 0
Subsite (Other and NOS)	0.05234	0.43
Diagnosed 1985-1994	0.21527	≈ 0
Respiratory issues	0.03873	0.41

Table 7: A summary of the variables in the log-normal model.

This model had a log-likelihood of -49947.1, a concordance of 0.725, and an AIC score of 99322.07. Again, we decided to remove the 'resp' variable due to its high p-value, but again the model did not improve, with the new model having a log-likelihood of -49947.1, a concordance of 0.7249, and an AIC score of 99916.18. Again, the initial model was the best one to use.

2.3.5 Developing the log-logistic model

Finally, we fitted the following log-logistic model:

Variable	Coefficient	P-value
Intercept	148.09559	≈ 0
Sex (Male)	-6.33154	≈ 0
Age	-1.40663	≈ 0
Stage (Localised)	36.08219	≈ 0
Stage (Regional)	1.82534	0.3029
Stage (Distant)	-35.50103	≈ 0
Subsite (Coecum and ascending)	5.10127	≈ 0
Subsite (Descending and sigmoid)	3.67464	0.003
Subsite (Other and NOS)	6.31956	0.0017
Diagnosed 1985-1994	-2.44746	0.006
Respiratory issues	0.10107	0.9443

Table 8: A summary of the variables in the log-logistic model.

This had a log-likelihood of -61885.6, concordance of 0.7331, and an AIC score of 123795.1. Removing the resp variable gives a log-likelihood of -62274, a concordance of 0.733, and an AIC score of 124570.1, so we again kept the saturated model.

2.4 The Random survival forest model

As explained in Section 1.2.2, a key part of the random survival forest (RSF) model is splitting up the data to form many decision trees. We used two methods to do this: the log-rank splitting method and log-rank-score splitting method. These methods ensure that the partitioning process considers both the censored observations and the uncensored observation when determining the optimal splits in the data. They work by evaluating the differences in survival probabilities between groups based on the observed event times and the censored survival times, allowing for a more comprehensive partitioning of the data. Moreover, they are inherently robust to censoring because they focus on comparing survival experiences across groups rather than relying solely on observed event times. This robustness allows them to handle varying degrees of censoring in survival data without compromising predictive accuracy.

2.4.1 The log-rank splitting rule

At each node of a tree within the forest, the dataset is split based on all possible values for a given covariate. The log-rank test is applied to compare the survival curves of the two groups resulting from each potential split. The split that yields the highest log-rank test statistic (indicating the most significant difference in survival) is chosen to split the node. This process is then repeated recursively.

Mathematically, the splitting value is calculated using log-rank test given as:

$$L\left(X_j, c\right) = \frac{\sum_{k=1}^{K}\left(d_{k,l} - Y_{k,l}\frac{d_k}{Y_k}\right)}{\sqrt{\sum_{k=1}^{K}\frac{Y_{k,l}}{Y_k}\left(1-\frac{Y_{k,l}}{Y_k}\right)\left(\frac{Y_k-d_k}{Y_k-1}\right)d_k}} \tag{3}$$

The larger the absolute value of $L(X_j, c)$, the greater the survival difference between the two groups. The best split is determined by finding the predictor X_j* and split value c* which maximises the value of this statistic [8].

2.4.2 The log-rank score splitting rule

For this splitting method, we assume the variable X has been ordered so that $X_1 \leq X_2 \leq \cdots \leq X_n$. Now we compute the 'ranks' for each survival time T_j,

$$a_j = \delta_j - \sum_{k=1}^{\Gamma_j}\frac{\delta_k}{n-\Gamma_k+1} \tag{4}$$

where $\Gamma_k = \#\{t : T_t \leq T_k\}$,The log-rank score test is defined as

$$S(x,c) = \frac{\sum_{X_j \leq c} a_j - n_L\bar{a}}{\sqrt{n_L\left(1-\frac{n_L}{n}\right)s_a^2}} \tag{5}$$

where $\bar{a}$ and s_a^2 are the sample mean and sample variance of $\{a_j : j = 1, \ldots, n\}$ and $n = n_L + n_R$, where n_L is the sample size of the left daughter node. The log-rank score splitting

method defines the measure of node separation as $|S(X, c)|$. Maximising this value over X and c yields the best split [6].

2.4.3 Random survival forest model analysis

We fitted two RSF models, including survival trees built using log-rank and the log-rank-score splitting rules on the dataset. These two models were built using 80% of our total dataset as our training dataset, and the remaining 20% as our testing dataset. The characteristics of the two fitted models are summarized in Table 9 below.

Characteristic	**Log-Rank**	**Log-Rank-Score**
Sample size	12450	12450
Number of trees	100	100
Forest terminal node size	15	15
Average number of terminal nodes	477.53	473.09
Number of variables tried at each split	3	3
Total number of variables	5	5
Family	surv	surv
Number of random split points	10	10
Mean Square Error	2451.399	2450.87

Table 9: Random survival forests results using log-rank and log-rank-score splitting rules.

We fitted two RSF models, each with 100 survival trees but built using the log-rank and log-rank score splitting methods respectively. The results of the mean square errors of each model are presented in Figure 13.

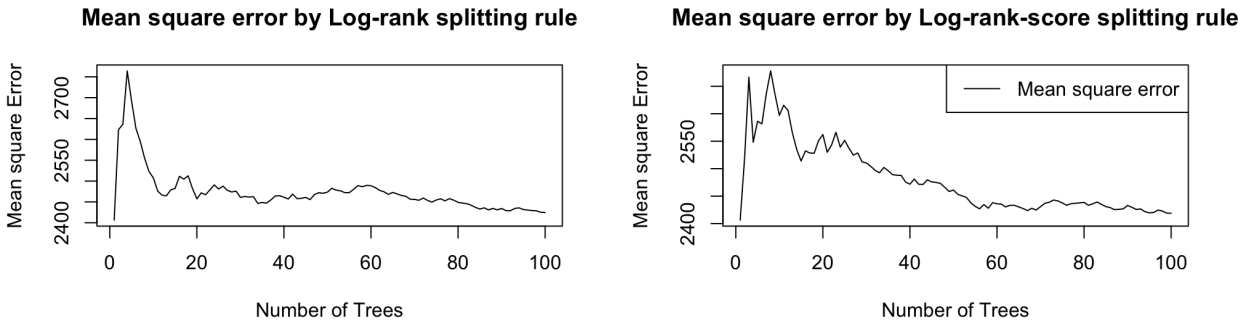
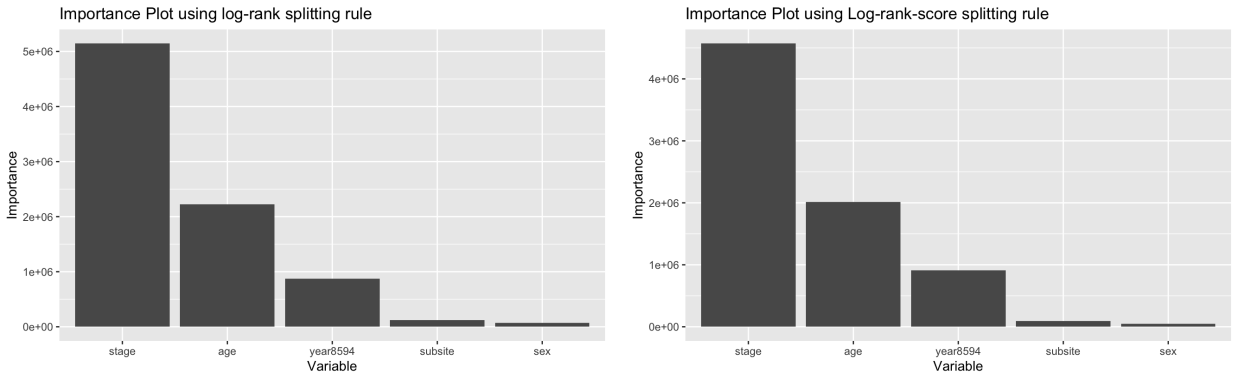


Figure 13: Plots showing the number of trees against the mean square error using the each of the two splitting rules.

Table 9 and Figure 13 show that the log-rank splitting rule is more stable than the log-rank score splitting rule. As Figure 13 shows, the log-rank splitting method becomes stable when the number of trees reaches about 40. However, log-rank-score splitting method becomes stable when the number of trees reaches about 60. When the number of trees reaches

maximum value which is 100, the mean square error for log-rank score splitting method is 2450.87, which performs better than log-rank splitting method for 2451.399. It make sense since the log-rank score method adding a score term, which needs longer time to fit the value, and eventually, predict the model in a more accurate way.

Since the unique decision pattern for random forest model, we can get each variable's importance values, hence, getting the most important variable from our model. The importance plots under log-rank splitting rule and log-rank score splitting rule are presented in Figure 14.



(a) Variable importance plot using the log-rank splitting rule

(b) Variable importance plot using the log-rank-score splitting rule

Figure 14: Variable importance plots for both splitting rules

As we can see from Figure 14, both splitting methods gave the same results for the importance rank of each variables. The variable 'stage' is found to be the most important variable, with an importance level of 5030015.38 for the log-rank splitting method, and 4872162.30 for the log-rank score splitting method. The variables 'age', 'year8594' and 'subsite' are decreasingly important. And for the variable 'sex', the importance level was the lowest - 62763.28 for log-rank splitting method and 62811.09 for the log-rank score splitting method [10].

3 Model evaluation and further discussion

3.1 Evaluation of each model

3.1.1 The Cox proportional hazards model

Strengths:

- **Flexibility:** No need to choose a particular probability distribution in advance. Besides, even though the baseline hazard in the model remains unspecified, it is still possible to produce parameter estimates to evaluate the effect of multiple explanatory variables.

- **Considering Censored Cases:** Cox models consider both the likelihood of an event outcome and the likelihood estimator including the information contained in the survival times of censored cases.
- **Versatility:** It is easy to consider time-varying covariates to adapt to actual situations.

Weaknesses:

- **Difficult to Satisfy Assumptions:** The proportional hazards assumption is often not realistic.

3.1.2 The Weibull model

Strengths:

- **Flexible (as a parametric model):** Although the Cox model is more flexible, as a parametric model, the Weibull model is more flexible than alternatives, due to the fact it takes in parameters for both the shape and the scale of the hazard.
- **Easy to Check Assumptions:** As shown in Section 2.3.1, the assumptions for whether a Weibull distribution can be fitted can be satisfied by a simple graphical check.

Weaknesses:

- **Parametric Model Inflexibility:** Although it is flexible compared to other parametric models, compared to the Cox model it is less flexible, due to the need to assume the hazard function is monotonic.

3.1.3 The log-normal and log-logistic models

Strengths:

- **Non-monotonic Hazard Functions:** These two models do not require the assumption that the hazard is monotonic, so the hazard can go up and down as time progresses, making it useful for certain data sets.

Weaknesses:

- **Hard to Fit:** The strength of the model is also a drawback, as it will not always be a good idea to fit this model to all types of data.

3.1.4 The random survival forest model

Strengths:

- **High Prediction Accuracy:** Random forest models can handle complex datasets with nonlinear relationships and interactions among variables, leading to potentially high prediction accuracy.
- **Robustness to Overfitting:** Random forests mitigate overfitting by constructing multiple decision trees and averaging their predictions, making them less sensitive to noise and outliers in the data.

Weaknesses:

- **Computationally Intensive:** Training a random forest model can be computationally intensive, especially for large datasets with many features and observations. It took approximately ten minutes for us to obtain our final results in Section 2.4.3.
- **Can be difficult to interpret:** While random forest models offer variable importance measures, interpreting the precise relationship between predictors and outcomes may be challenging due to the complexity of the model.

3.2 Model comparison

Finally, we compare our 7 models: the Cox proportional hazards model, the Cox time-varying proportional hazards model, the Weibull model, the log-normal model, log-logistic model and our two RSF models. We use the concordance score for each model as the primary measure to do this. In this part, we take 4/5 data as our train data, 1/5 data as our test data. Thus, there will exist little inconsistency from result here and above.

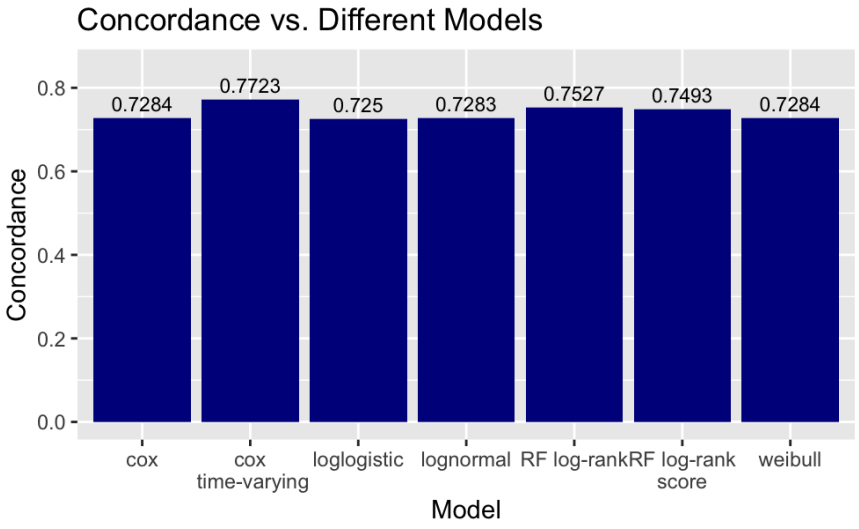


Figure 15: A bar chart showing the concordance scores for each of our 7 models

The concordance scores shown in Figure 15 are very similar for all models, with the largest difference between any two given models being less than 0.05. The highest concordance score is for the Cox time-varying proportional hazards model at approximately 0.77. The concordance scores for the machine learning models are overall slightly higher than non-machine learning models, each with a concordance of about 0.75. Overall, all of our models demonstrated comparable accuracy and have very similar predictive power. Thus, it is difficult to pinpoint one model as being the outright 'best' model for our data.

4 Conclusions

4.1 Main results

In the Cox proportional hazards model, we use stratification and time-varying methods dividing data in order to satisfy the Cox proportional hazards assumption. After improving our cox model, we get a pretty high concordance level for about 0.77 and summarize that the variables 'age', 'subsite', 'stage' and 'year8594' are quite important for our model. In the parametric models, we tried Weibull distribution, Lognormal distribution and Log-logistic distribution, after testing initial assumptions, we firstly figure out the concordance for 3 models are all approximately 0.73 and try to find out each variable's p-value to get an idea whether certain variable is important. From these models, we identified that the results are slightly different from what we found for the most important variable in Cox model, 'sex', 'age', 'subsite' and 'stage' and 'year8594' as being the variables with the highest influence on patients' survival outcomes now.

Afterwards, the random survival forest model was applied using the log-rank and log-rank score splitting rules. 100 trees were used to make the decisions for each of these methods. We then compared the two splitting rules by plotting the mean square error against the number of trees. As a result, the log-rank splitting rule is more stable than log-rank score splitting rule. Moreover, our RSF models found 'stage' to be the most influential variable, with the variables 'age', 'year8594' and 'subsite' also having significant predictive power.

After constructing all 7 models, we used the concordance score for each to compare between them. These results showed there was not a large difference between each model. Thus, it is difficult to pinpoint one model as our 'best' model.

4.2 Comparison with published work

There are considerable researches into colon cancer and the variables that affect a patient's prognosis. Cancer Research is the main source for a lot of our statistics as they are a charity that receive a substantial amount of donations to research all types of cancer. Cancer Research have published information on 'What affects survival of colon cancer?' [2]. Their research uncovered that the stage of the cancer greatly affects a patients prognosis. Through our importance plot under random survival forest model, we also find out that the variable 'stage' as our most important variable.

Furthermore, from the paper entitled 'Predictors of colorectal cancer survival using cox regression and random survival forests models', it is apparent that the Cox model had a better predictive performance [10]. However, in that paper, all the covariates satisfy the proportional hazards assumption. When we set up our Cox model, we also found that the concordance was quite low when the assumption is not satisfied. Thus, we tried to use stratification and time-varying methods to satisfy the Cox model's assumption. After improving our model, the concordance significantly increased, which reaches a consensus with the paper. Moreover, this paper found that the RSF model did not pick the variable 'sex' as an important variable, while it is significant in the Cox model. However, in our data, using the Cox model, we did not pick out the variable 'sex' as an important variable. This can be explained by the fact we are using a different datasets and different p-values to decide whether to reject hypotheses.

4.3 Reflection and further improvement

While our existing models proficiently forecast survival time, there exists ample room for refinement to enhance accuracy.

Firstly, the random survival forest models we fitted could easily be improved by increasing the number of trees used in order to reduce errors and increase accuracy. For us, though, it was too computationally expensive to use more than 100 trees.

Additionally, our preliminary investigations suggest that the variables initially provided for model development do not fully encapsulate all factors influencing a patient's prognosis. Elements such as lifestyle, dietary habits, ethnic background, familial medical history, alcohol consumption, and numerous other factors undoubtedly exert significant influence on survival outcomes. We are also not provided with information on the patient treatment plan. We can assume this would be essential when devising a model. To enhance our model's efficiency, it is imperative that we incorporate this additional information. Further research should build on this work by seeking and incorporating these extra factors into a model.

As well as this, if we wanted to improve our models to make them more applicable, the data could be expanded outside of just one northern European country to apply to more countries around the world. The study could also be expanded, not just to limit censored data, but also to help apply findings to modern patients.

References

[1] American Cancer Society. Colon and rectal cancer survival rates, 2023. Accessed on DATE.

[2] Cancer Research UK. Bowel cancer survival. `https://www.cancerresearchuk.org/about-cancer/bowel-cancer/survival`, 2023.

[3] Sachin Date. The akaike information criterion. `https://timeseriesreasoning.com/contents/akaike-information-criterion/`, 2022.

[4] Steven J Durning, Anthony R Jr Artino, Thomas J Beckman, John Graner, Cees van der Vleuten, Eric Holmboe, and Lambert Schuwirth. Enhancing the quality of case reports in health professions education: Key features and guidelines. Academic Medicine, 88(11):1472–1481, 2013.

[5] Soraya et. al. Evaluation of prognostic factors effect on survival time in patients with colorectal cancer, based on weibull competing-risks model. Scientific Reports, 2017.

[6] Hemant Ishwaran, Michael S Lauer, Eugene H Blackstone, Min Lu, and Udaya B Kogalur. Randomforestsrc: Random survival forests vignette. Random Forest SRC, 15, 2022.

[7] Inhwa Lee, Seung-Hyun Baek, Hyunsung Kim, Hong-Jae Jo, Nahm-Gun Oh, and Sanghwa Ko. Survival analysis for colon subsite and rectal cancers: Experience from a single surgeon. Korean Journal of Clinical Oncology, 11(2):114–119, 2015. Published online December 31, 2015.

[8] Yingxin Liu, Shiyu Zhou, Hongxia Wei, and Shengli An. A comparative study of forest methods for time-to-event data: variable selection and predictive performance. BMC Medical Research Methodology, 21:1–16, 2021.

[9] Melinda Mills. Introducing Survival and Event History Analysis. SAGE Publications Ltd, 2011.

[10] Mohanad Mohammed, Innocent B Mboya, Henry Mwambi, Murtada K Elbashir, and Bernard Omolo. Predictors of colorectal cancer survival using cox regression and random survival forests models based on gene expression data. PLoS One, 16(12):e0261625, 2021.

[11] Steven J Rigatti. Random forest. Journal of Insurance Medicine, 47(1):31–39, 2017.