## Chapter 1

## Introduction and Objectives

### **Introduction**

Head and neck cancer (HNC) refers to a group of malignant tumours that develop in or around the throat, larynx, nose, sinuses, and oral cavity. As per NCDIR and ICMR, the primary site of head and neck cancer includes the tongue, mouth, tonsil, oropharynx, nasopharynx, hypopharynx, pharynx, and larynx [1,2]. The majority of head and neck cancers, about 90%, are squamous cell carcinomas, which arise from the flat, thin squamous cells lining the inner surfaces of the mouth, nose, and throat [3]. Tobacco and alcohol consumption are the leading risk factors for HNC, with their combined use significantly amplifying the risk due to a synergistic effect that accelerates cancerous cell development. In addition to these, other risk factors include exposure to human papillomavirus (HPV), particularly in oropharyngeal cancers, environmental exposure to harmful chemicals, pollutants, and radiation, and a family history of cancer [4,5]. The treatment for head and neck cancer (HNC) depends on factors like the cancer's location, size, type of cells, and the patient's overall health. Common treatments include surgery, radiation therapy, and chemotherapy. Treatment plans are personalized to optimize the patient’s health outcomes. Early detection and treatment are essential for improving survival rates, underscoring the importance of awareness regarding symptoms and risk factors for timely diagnosis and intervention [6].

Head and neck cancers rank 6th globally in terms of incidence, accounting for about 4.75% of all cancer cases. In 2022, an estimated 947,211 cases of HNC were reported worldwide, out of 19.98 million cancer cases, according to GLOBOCAN estimates. These cancers include lip and oral cavity malignancies, salivary glands, oropharynx, nasopharynx, hypopharynx, and larynx. Among these, cancers of the lip and oral cavity were the most common, with 389,846 cases, followed by laryngeal cancers (189,191 cases) and nasopharyngeal cancers (120,434 cases). This distribution highlights the diversity of head and neck cancers and the geographical differences in their prevalence [7]. According to the Hospital-Based Cancer Registry (HBCR), 2022 report from the Malabar Cancer Centre, head and neck cancers represented 14.39% of all reported malignancies, with 872 cases out of a total of 6,059 cancers, making it the second most common type after digestive organ malignancies. Among these cases, 76.26% (665) were males and 23.74% (207) were females, indicating a significantly higher prevalence in males. Lifestyle factors play a major role, with smoking reported in 32.8% of cases, alcohol consumption in 29.9%, pan chewing in 28.4%, and a family history of cancer in 19.3%. These statistics highlight the importance of addressing preventable risk factors to mitigate the burden of head and neck cancers [8].

Head and neck cancer (HNC) is a major public health concern in India, with it accounting for 26% of all cancer cases in males and 8% in females. India experiences a higher burden of HNC compared to countries like the USA, UK, and Brazil. The northeastern regions report the highest incidence of HNC, with cancers of the mouth being the most common, followed by cancers of the tongue, larynx, hypopharynx, and tonsils. The burden is particularly higher in those aged 60 and above, compared to younger age groups [9]. The incidence and mortality rates of head and neck cancers (HNC) are significantly influenced by demographic factors such as age and gender. Older populations, particularly those aged 60 and above, have higher rates of HNC due to the cumulative effects of risk factors and biological aging. Gender differences are also notable, as men generally have a higher incidence of HNC compared to women. This discrepancy is likely attributed to higher exposure to risk factors such as tobacco and alcohol use, which are more prevalent in males. Clinical features, including the tumor stage at diagnosis and the involvement of lymph nodes, are crucial in determining survival outcomes. Tumors detected at more advanced stages, with larger sizes and greater nodal involvement, tend to have poorer prognoses and higher mortality rates, as these characteristics often indicate more aggressive cancer progression. Early detection and staging play a pivotal role in improving survival rates, making it essential to address these clinical factors in managing HNC [10].

Survival analysis is a statistical method commonly used in cancer research to assess the time to an event of interest, such as recurrence, progression, or death. Key methodologies include the Kaplan–Meier (KM) method, log-rank tests, and the Cox proportional hazards regression model. These statistical tools are essential for understanding survival trends, comparing treatment outcomes, and identifying critical prognostic factors in cancer research [11]. Various factors influencing survival outcomes for head and neck cancer patients include demographic characteristics, clinical features, and treatment modalities. Despite advancements in therapy, survival remains a significant concern due to the high risk of recurrence and second primary malignancies, particularly in the head and neck, lungs, and esophagus.

The Kaplan-Meier estimator was created by Edward L. Kaplan and Paul Meier in 1958. It is a non-parametric statistical method widely used to estimate survival functions and analyze time-to-event data. This method provides an insightful graphical representation of survival probabilities over time and effectively handles censored data. It has become a standard approach in clinical and epidemiological research. Kaplan-Meier survival curves help identify differences in survival among various groups based on factors such as age, gender, and tumour stage. This makes it an essential tool for comparing patient outcomes, such as survival rates after treatment and the effectiveness of interventions in clinical trials. By estimating the fraction of individuals who survive specific events such as death or disease recurrence, the Kaplan-Meier method remains a foundational aspect of survival analysis, particularly in studies involving incomplete follow-up data [12].

The Cox proportional hazards model is a widely utilized statistical tool in survival analysis, aimed at exploring the relationship between the time until an event of interest occurs commonly death or failure, and one or more predictor variables. This regression model provides a method to assess how various clinical and demographic factors influence survival times while adhering to the crucial assumption that the effect of these predictor variables on the risk of the event occurring (hazard) remains constant throughout the study. In constructing a Cox model, it is essential to carefully choose relevant predictor variables that will be included in the analysis. To achieve this, researchers often employ variable selection techniques, such as forward selection, which incrementally adds predictors based on their significance, and backward elimination, which starts with all potential predictors and sequentially removes the least significant ones. These methods make the model simpler and help focus on the most important factors. Furthermore, to ensure a balance between model fit and complexity, the Akaike Information Criterion (AIC) is frequently used. The AIC provides a statistical measure that penalizes excessive complexity in the model; therefore, it guides researchers in identifying models that briefly explain the data while avoiding overfitting. By prioritizing only, the most relevant variables based on the AIC, analysts can develop a robust Cox model that accurately reflects the underlying survival data, ultimately leading to more insightful conclusions about the factors influencing survival outcomes [13].

Lasso regression, which stands for Least Absolute Shrinkage and Selection Operator, is a valuable method used to manage high-dimensional datasets that have a large number of predictor variables. In scenarios where the number of variables exceeds the number of observations or when multicollinearity among predictors is present, traditional regression techniques may struggle to produce reliable results. The key feature of Lasso regression is its application of a penalty term to the ordinary least squares regression model. This penalty is proportional to the absolute values of the coefficients associated with each predictor. By imposing this constraint, Lasso effectively shrinks the coefficients of less important variables towards zero. As a result, it eliminates unimportant features from the model, effectively conducting variable selection. This dual functionality not only simplifies the model by retaining only the most relevant variables but also helps to prevent overfitting, thereby enhancing the model's accuracy and predictive power [14].

The C-index, also known as the concordance index, and the time-dependent area under the curve (AUC) are two important metrics used to evaluate the accuracy of predictive models in survival analysis. The C-index measures the model's ability to correctly rank predictions, indicating how well it distinguishes between survival times. On the other hand, the time-dependent AUC quantifies the model's performance over time by calculating the area under the time-dependent receiver operating characteristic (ROC) curve, which plots the true positive rate against the false positive rate at various time points. Together, these metrics provide valuable insights into the reliability and effectiveness of a model's survival predictions over time.

This research study is designed to thoroughly apply various survival analysis techniques, specifically the Kaplan-Meier estimator, the Cox proportional hazards model, and the Lasso-based Cox regression model and to systematically identify significant prognostic factors that influence the outcomes of patients suffering from head and neck cancer who are treated at the Malabar Cancer Centre. By applying these advanced methods, we aim to deepen our understanding of the survival patterns among head and neck cancer patients.

### **1.2 Objectives**

1. To identify significant prognostic factors influencing survival in head and neck cancer patients using Cox proportional hazards regression model and Lasso-based Cox model.
2. To estimate the 5-year survival probability and 95% confidence intervals for head and neck cancer patients based on key prognostic factors.
3. To assess the improvements in survival prediction using Lasso-based Cox regression over traditional Cox regression for head and neck cancer

## Chapter 2

## Review of Literature

1. **Dongdong Zhou et al. (2021)** conducted a study in which they applied Lasso Cox regression to improve the prediction of prognosis for patients with AFP-negative liver cancer (AFP-NHCC). In their research, they developed Nomogram1 using Lasso Cox regression, which achieved a C-index of 0.708 and an AUC (Area Under the Curve) of 0.736. Notably, Nomogram1 outperformed Nomogram2, which utilized the traditional Forward Stepwise Cox method, with the latter yielding a C-index of 0.706 and an AUC of 0.714. Furthermore, in a validation cohort of patients—separate from the original dataset used to develop the models—Nomogram1 exhibited improved performance, achieving a C-index of 0.752 and an AUC of 0.784. These enhanced metrics signify the robustness and reliability of Nomogram1 in predicting patient outcomes for those with AFP-NHCC. The results from this study highlight the significant advantages of employing Lasso-Cox regression over traditional modeling methods. By effectively identifying and concentrating on the most important clinical factors, this innovative approach not only boosts the accuracy of prognosis predictions but also offers a valuable strategy for tailoring treatments to the individual needs of AFP-NHCC patients [15].
2. **Yuxin Xu et al. (2021)** conducted a study aimed at developing a prognostic model to predict overall survival (OS) in patients diagnosed with synchronous colorectal carcinoma (SCC). The research utilized a substantial dataset that included 4,616 patients, from which the authors initially considered 31 potential variables related to survival outcomes. By applying the LASSO method, the study effectively narrowed down these variables to 11 key predictors deemed most relevant for the prognosis of SCC patients. This process involved systematically evaluating the influence of each variable, leading to the exclusion of those that were less significant in the model. When comparing the LASSO-based model to traditional Cox regression models, the authors found that, while performance metrics were generally similar, the LASSO model outperformed the Cox model in time-dependent areas under the curve (time AUC). This finding underscored its enhanced capability for making precise predictions regarding long-term survival outcomes, which is critical for treatment planning and patient counselling. Notably, the study revealed that certain variables, such as tumor grade and marital status, were excluded from the LASSO-based Cox regression model due to issues related to multicollinearity and the potential for overfitting. Nevertheless, both models retained common predictors, including age, sex, tumor size, and cancer stage, as indicated by pT, pN, and pM classifications. To assess the robustness and reliability of the models, both were subjected to validation processes. The findings indicated that the LASSO-based model consistently outperformed its traditional counterpart, providing a greater net benefit in clinical decision-making. Its improved efficiency and accuracy make it a more viable tool for clinicians aiming to predict the prognosis of patients diagnosed with synchronous colorectal carcinoma [16].
3. **Lu Bhai and Daniel Gillen (2017**): This paper examines survival analysis using Cox proportional hazards additive models, which extend the traditional Cox model by incorporating bivariate additive modeling. While the Cox model is commonly used to assess the relationships between predictors and mortality risk for censored survival data, it depends on a parametric assumption of linearity between covariates and mortality risk. In contrast, generalized additive models (GAMs) offer a more flexible framework that allows for nonlinear effects of predictors while preserving additivity. By integrating bivariate additive modeling into the Cox framework, this study enhances the analysis of complex relationships between predictors and mortality risk. The primary focus is on estimating geolocation effects in spatial epidemiologic studies, addressing the challenges of modeling survival data that are influenced by spatial variability. This approach provides a robust method for understanding spatial factors in survival analysis and contributes valuable insights to epidemiologic research [17].
4. **Sijin Sun et al. (2020)** conducted a comprehensive study to develop a prognostic signature specifically for lung adenocarcinoma (LUAD) by leveraging gene expression data. The research began with the careful selection of 1,057 genes from two primary groups identified within the TCGA-LUAD dataset, a rich source of genomic information on lung cancer. To assess the relationship between these genes and patient survival outcomes, the team applied univariate Cox regression analysis. This statistical method allowed them to identify specific genes that were significantly associated with survival rates among LUAD patients. They used a stability selection method to identify key candidate genes that consistently demonstrated significant associations across tests, ensuring reliable findings. To refine their results, the researchers used a Lasso-penalized Cox regression model, which incorporates regularization to prevent overfitting while selecting predictive variables. Optimal parameters were determined through cross-validation, assessing the model's predictive accuracy by partitioning the dataset into training and validation sets. Through this careful method, they ultimately identified a robust set of genes that have strong predictive capabilities concerning the survival of patients diagnosed with LUAD. Their research enhances LUAD prognosis understanding and may aid in developing personalized treatment strategies [18].
5. **Qi Wang et al. (2022):** They investigated risk factors for recurrence in early-stage hepatocellular carcinoma (HCC) patients after minimally invasive treatments and developed a predictive model using Lasso-Cox regression. The study analyzed clinical data from 547 patients treated from January 2012 to December 2016. Key predictors identified through Lasso regression included age, gender, liver cirrhosis, tumor number, tumor size, the PALBI index, and viral load. These predictors were incorporated into a Cox proportional hazards regression model. Patients were randomly divided into training (n = 254) and validation (n = 255) sets. Over a median follow-up of 59.3 months, 397 patients recurred, and 189 died. Recurrence rates were 26.0% at 1 year, 57.8% at 3 years, and 68.2% at 5 years; overall survival rates were 98.9%, 86.8%, and 74.4%, respectively. The Kaplan-Meier analysis highlighted that recurrence significantly impacted overall survival. The Cox model showed strong predictive performance with 0.729 (training) and 0.726 (validation) C-indices. A nomogram based on this model was developed to help clinicians estimate individual patient recurrence risks, enabling personalized follow-up strategies to improve outcomes for early-stage HCC patients [19].
6. **Jie Zhou et al. (2024)** developed a predictive nomogram for assessing suicide risk in lymphoma patients by combining the Cox proportional hazards regression model with LASSO regression. The Cox model identified significant risk factors such as age, gender, ethnicity, marital status, cancer stage, and treatment options. LASSO regression refined these predictors by shrinking less relevant coefficients, addressing multicollinearity, and reducing overfitting. The resulting nomogram achieved a C-index of 0.773 in the training and 0.777 in the validation set, demonstrating excellent calibration and clinical utility. This integration of methods enhances predictive models for clinical decision-making [20].
7. **Xiao Jia et al. (2024):** Head and neck squamous cell carcinoma (HNSCC) is a common cancer affecting the mucosal epithelium of the oral cavity, pharynx, and larynx, and it poses a significant global health challenge. Despite improvements in treatment options like surgery, chemotherapy, radiotherapy, and CAR-T immunotherapy, patients with advanced-stage HNSCC face a poor prognosis, with about 50% experiencing unfavorable outcomes. To enhance targeted therapies and improve patient survival, the integration of prognostic signatures into clinical practice has gained attention. Metal ions, such as copper and iron, play essential roles in cellular regulation, but their dysregulation can lead to specific cell death forms, including cuproptosis and ferroptosis, both relevant in various cancers, including HNSCC. This study aimed to develop a prognostic model based on cuproptosis- and ferroptosis-related genes, comprising 12 key genes validated through rigorous statistical analyses. Notably, the gene AURKA was identified as critical for HNSCC progression, where its knockdown significantly hindered cell proliferation and migration. These findings suggest that integrating cuproptosis and ferroptosis signatures could improve prognostic assessments and therapeutic strategies in HNSCC, with AURKA as a promising biomarker and therapeutic target [21].
8. **Silei Sui et al. (2020)** conducted a study aimed at developing a prognostic model for breast cancer (BC) that utilizes tumor-infiltrating immune cells (TIICs). The research emphasizes the integration of Cox proportional hazards regression and the least absolute shrinkage and selection operator (LASSO) regression for identifying biomarkers. By analyzing transcriptomic data from 5,112 patients using CIBERSORT, the study quantified 22 types of TIICs. Univariate Cox regression identified 12 immune cell types significantly associated with overall survival (OS). LASSO regression further streamlined these findings into six key biomarkers used in an immune prognostic model, which effectively stratified patients into high- and low-risk groups with notably different survival outcomes. The immune score derived from this model demonstrated robust prognostic value across training, validation, and test cohorts, and it also improved predictions for chemotherapy responses. When combined with TNM staging in a nomogram, the approach enhanced predictive accuracy, highlighting the synergistic potential of integrating immune-based metrics with traditional prognostic assessments. These findings showcase the effectiveness of utilizing Cox and LASSO methods to leverage immune infiltration data for breast cancer prognosis and treatment planning [22].
9. **Hadi RaeisiShahraki, Alireza Salehi, and Najaf Zare (2015):** This study investigated survival factors in male breast cancer patients using the LASSO-Cox regression method, comparing it to the traditional Cox proportional hazards model. Based on data from 50 patients in Fars province, Iran (1989-2008), the analysis used multiple imputations to manage missing data. Survival rates were reported at three years (92%), five years (77%), and ten years (26%), with an average survival time of 62 months. Key prognostic factors included age at diagnosis, alcohol use, nipple discharge, tumor laterality, histological grade, and symptom duration. The LASSO-Cox method outperformed the Cox model, eliminating eight low-effect variables and reducing standard errors by 2.5 to 7 times, enhancing precision and efficiency, especially in handling multicollinearity. The optimal λ for the LASSO-Cox model was 0.018. While it confirmed prior associations of histological grade and tumor size with poorer survival, it also indicated that alcohol use and nipple discharge may correlate with increased survival, needing further research. Limitations included missing data on marital status, metastasis, and treatments, along with a single-center dataset. Overall, the study suggests that LASSO-Cox regression can improve model stability and reliability in small sample sizes, with potential broader applications in clinical and cancer research [23].
10. **Georg Hahn et al. (2024):** This study aims to improve how we predict a person's risk of disease. It goes beyond traditional risk models that only give basic estimates of risk. First, it focuses on predicting time-dependent hazards and survival rates for a clearer understanding of how long someone might remain disease-free. Second, it shows how to use LASSO-regularized Cox proportional hazards models for this task. The researchers used data from a genetic study on Alzheimer's disease that involved 6,792 participants (4,102 with the disease and 2,690 without) and considered 87 factors like sex and genetic information. The LASSO method helped handle a lot of data while including important factors related to health. The results showed that this approach produced more accurate survival predictions than traditional methods. It also allowed doctors to create personalized survival curves, giving a clearer view of how the disease may progress for individual patients. The method is fast and efficient, with personalized survival curve calculations taking less than a minute for the whole dataset. This shows that LASSO-based survival analysis can be practical and useful in healthcare for precision medicine and assessing risk [24].
11. **Xintian Cai et al. (2020):** This study aimed to create and confirm a prognostic model for the 5-year chance of developing Type 2 Diabetes (T2D) in high blood pressure patients in China. The research utilized LASSO (Least Absolute Shrinkage and Selection Operator) regression to identify the most significant determinants of T2D. Key aspects such as age, body mass index (BMI), fasting plasma glucose (FPG), and total cholesterol levels were found to be the most critical in anticipating the risk of T2D. The model was developed using a derivation cohort and validated in an independent cohort, showing strong capability. The area under the receiver operating characteristic curve (AUC) for the derivation cohort was 0.878, while the AUC for the validation cohort was 0.855. These results suggest that the model can accurately predict the probability of T2D in hypertensive patients, allowing for early intervention and better handling of at-risk individuals. The study underscores the importance of routine screening and the potential for predictive models in preventing diabetes, especially in groups with hypertension [25].
12. **Yi Wang et al. (2021):** This research developed an eight-gene signature to forecast outcomes in pediatric brain tumor (PBT) patients utilizing data from the Pediatric Brain Tumor Atlas: CBTTC cohort. The cohort was divided into training (486 patients) and validation (487 patients) sets. Key genes (CBX7, JADE2, IGF2BP3, OR2W6P, PRAME, TICRR, KIF4A, and PIMREG) were identified and refined through LASSO Cox regression, revealing significant survival disparities between high- and low-risk groups. The risk score demonstrated strong predictive ability and was independent of other clinical factors. Furthermore, analyses suggested interactions with immune-related pathways, indicating potential for personalized prognosis and innovative therapeutic strategies in PBT patients [26].
13. [**Coralie R Arends**](https://pubmed.ncbi.nlm.nih.gov/?term=Arends+CR&cauthor_id=31693181) **et al. (2020):** This study developed and validated a clinical prediction model (CPM) for survival in hypopharynx cancer to enhance personalized survival estimates. A retrospective analysis of 768 patients was performed, with data split into derivation and validation sets. The final CPM included factors like gender, TNM classification, ACE27 score, BMI, hemoglobin, albumin, and leukocyte count, with ACE27, BMI, hemoglobin, and albumin identified as key predictors. The model achieved a C statistic of 0.62, indicating its ability to distinguish between risk groups and outperforming a TNM-only model. The study highlights the importance of considering patient-specific factors like comorbidity and nutritional status in predicting survival, while further research is needed to refine the model [27].
14. **Shiyan Li et al. (2021):** In this study univariate Cox regression, LASSO regression, and Cox proportional hazards regression were employed to identify key genes that impact the prognosis of cervical cancer. Initially, univariate Cox regression was used to identify genes associated with survival, highlighting how each gene influences patient outcomes. Subsequently, LASSO regression was applied to refine the model by retaining only the most significant genes. This step reduced the risk of overfitting and enhanced the model's reliability. Following this, Cox proportional hazards regression focused on the important genes selected by LASSO, ultimately identifying six key genes: SLC25A5, ENO1, ANLN, RIBC2, PTTG1, and MCM5, which showed a strong correlation with cervical cancer survival. Among these genes SLC25A5, ANLN, RIBC2, and PTTG1 were identified as independent prognostic markers. By integrating univariate Cox, LASSO, and Cox regression analyses, the study developed a reliable risk model to help predict patient outcomes and inform treatment decisions for cervical cancer [28].

## Chapter 3

## Methodology

### **3.1 Survival Analysis**

Survival analysis is an important set of statistical methods used to study time-to-event data, focusing on the duration until a significant event occurs, such as the failure of mechanical components or the death of living organisms. Although it primarily deals with time-related events, it can also be applied to other outcomes like skill development and financial performance. The origins of survival analysis lie in the study of mortality tables, with its early developments closely linked to World War II, where it played a vital role in evaluating the reliability (durability) of military equipment. Initially, the emphasis was on parametric models suited for engineering applications. However, in recent eras, there has been a notable shift towards non-parametric methods. This transition has been driven by the demands of medical research and clinical trials that require more flexible modeling techniques. Nowadays advancements in non-parametric techniques have become central to survival analysis, particularly in medical fields such as cancer prognosis and other health-related studies. This transition highlights the critical role of survival analysis in providing reliable insights and supporting informed decision-making across various domains.

#### **3.2 Survival Function and Hazard Rate**

The survival function is a basic concept in survival analysis. It is a statistical measure that represents the probability that a specific event of interest such as death, failure, or another outcome has not occurred by a certain point in time. It indicates the likelihood that an individual or object will survive beyond a given time point t. It is denoted as S(t) and is defined by:

Let T represent the time until the event occurs, and let t denote a specific point in time. The cumulative distribution function (CDF), denoted as F(t), indicates the probability that the event has occurred by time t. It is related to the survival function by,

The probability density function (PDF), f(t), describes the likelihood of an event occurring at an exact time t and is given by,

This means that it is the negative derivative of the survival function concerning time.

The hazard function is a crucial concept in survival analysis that measures the rate at which an event occurs at a specific time, given that the individual or system has survived until that moment. It is typically denoted as h(t) and can be calculated by taking the ratio of the probability density function (PDF) to the survival function. It is given by,

The survival function can be derived from the hazard function as,

The hazard rate has the interpretation,

The cumulative hazard function is given by,

#### **3.3 Censoring**

Censoring is an important idea in survival analysis, which looks at the time until a specific event happens. It occurs when we don’t have complete information about the time of the event. This can occur if the event hasn’t happened by the end of the study, the individual withdraws from the study, or they are lost to follow-up.

#### **3.4 Types of Censoring**

Major types of censoring are type I censoring, type II censoring, and random censoring.

##### **3.4.1 Type I Censoring**

Type 1 Censoring occurs when a study is designed to end at a specified time set by the researcher. Any subject who does not experience the event by the end of this study period is considered censored. Mathematically, Type 1 censoring implies that each subject has a fixed censoring time Ci > 0. The event time Ti is observed if , otherwise, we only know that . It is one of the most common forms of censoring in survival analysis.

##### **3.4.2 Type II Censoring**

Type 2 Censoring occurs when a study is designed to end after observing a predetermined number of events (e.g., deaths or failures). Mathematically, in Type 2 censoring, only the r smallest event times in a random sample of size n are observed, where r is a specified integer between 1 and n. This scenario arises when all individuals (n) begin the study simultaneously, and the study ends once r (r<n) events have been recorded. Subjects who do not experience the event by the time rth event occurs are considered censored, meaning their exact event times are unknown but are known to exceed the rth observed time.

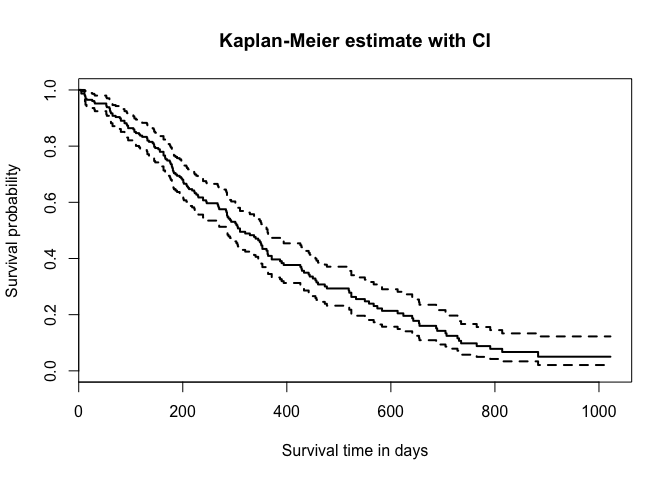
##### **3.4.3 Random Censoring**

Random Censoring occurs in studies where the observation period is fixed, but subjects join the study at different time points. During the study, some subjects experience the event of interest, others do not, some are lost to follow-up, and others remain event-free at the end of the study.

#### **3.5 Kaplan Meier Estimator**

The Kaplan-Meier estimate, also known as the product-limit estimate, is a non-parametric method used to estimate the survival function, particularly when dealing with censored data. It calculates the probability of survival at different time points by taking into account both observed events and censored observations. In this method, each subject's survival time is denoted as ​, with a corresponding status indicator or censoring indicator ​, where if the event occurred and if the data is censored. The Kaplan-Meier estimator is computed by multiplying the probabilities of surviving at each time point where an event occurs, adjusting for censoring. The formula for the estimator is given by,

where ​ represents the number of events at the time ​ and represents the number of individuals at risk just before time ​.



3.1 Kaplan Meier Survival Plot

The Kaplan-Meier estimator is visually represented as a step-like graph, where each step corresponds to an event occurring at a specific time. The graph does not show the right continuity indicators, as it typically jumps at times when events occur. Confidence intervals can be added around the survival curve. Right censoring is effectively depicted by marking censored data points with a small vertical tick mark or a symbol like a cross or a dot at the time when a subject is censored. This visualization makes it easier for researchers and clinicians to interpret and compare survival outcomes.

#### **3.6 Log-rank Test**

The Log-rank test is a statistical method used to compare the survival distributions of two or more groups, often in non-parametric models such as the Kaplan-Meier estimator. This test checks if there is a significant difference in survival between the groups over time, taking into account censored data, where some individuals may not experience the event or are lost to follow-up. The test statistic is given by,

Where, ​ is the observed number of events and ​ is the expected number of events. This statistic follows a chi-square distribution with degrees of freedom equal to the number of groups minus one.

The null hypothesis (​) assumes no difference in survival between the groups, while the alternative hypothesis (​) suggests that a significant difference exists.

* : There is no difference in survival distributions.
* ​: There is a significant difference in survival distributions.

If the p-value is less than 0.05 indicates that​ should be rejected in favour of ​, meaning a significant difference in survival exists.

#### **3.7 Cox Proportional Hazard (Cox PH) Model**

The Cox Proportional Hazards (Cox PH) model is a popular statistical method in survival analysis used to study the relationship between survival time and one or more predictor variables or covariates. It is considered a semi-parametric model and is commonly applied to time-to-event data, such as the time until death, disease progression, or system failure. This model is especially useful in multivariate survival analysis, where it helps assess the impact of multiple covariates on survival time. The Cox PH model is frequently used in fields like clinical research and epidemiology to handle complex data. The model assumes that the effect of covariates on the risk of an event (hazard) remains constant over time. In other words, the ratio of hazards between different groups stays the same throughout the study, which is known as the proportional hazard assumption. The Cox PH model can be expressed as,

Where,

* is the hazard at time t given the values of the covariates .
* is the baseline hazard function at time t represents the hazard when all covariates are zero.
* ​ are the coefficients for the covariates.
* ​ are the predictor variables (covariates).

##### **3.7.1 Hazard Ratio**

The quantities are referred to as hazard ratios, which are widely used in survival analysis to compare the hazard rates between two groups. The hazard ratio quantifies the impact of a covariate on the risk of the event of interest. Specifically, it represents the ratio of the hazard rates at different levels of a covariate.

* A hazard ratio of 1 indicates no difference in hazard between the groups.
* A hazard ratio greater than 1 suggests an increased hazard (higher risk) for the event.
* A hazard ratio less than 1 signifies a decreased hazard (lower risk) for the event.

##### **3.7.2 Estimation of Parameters of Cox PH Model**

The standard estimation method for the estimation of regression coefficients in the Cox proportional hazards model is by partial likelihood function and it can be represented as,

Where,

* is the set of observations at risk at time ti.
* is the status indicator for the ith individual.

The maximum partial likelihood estimates of is and is obtained by maximizing the partial log-likelihood function .

By differentiating with respect to and equate to zero subject to the condition that the second derivative of gives the MLE.

##### **3.7.3 Methods for Testing PH Assumption**

The proportional hazards (PH) assumption is a key principle in the Cox proportional hazards model, stating that the hazard ratios between groups remain constant over time. This means the effect of each covariate on the hazard rate is assumed to be time-invariant. If the PH assumption is violated, the model becomes unreliable, as the hazard ratios will no longer accurately reflect the effect of covariates over time. This can lead to biased results, incorrect conclusions, and misleading survival analysis. Grambch-Therneau test and Schoenfeld residual test are widely used methods for testing the proportional hazard assumption.

Grambch, Therneau, and Fleming proposed a global test based on the absolute value of summed Schoenfeld residuals in 1990. Later in 1994, Grambch and Therneau modified this test using scaled Schoenfeld residuals. The test statistic for Grambch and Therneau non-proportionality is given by,

Where,

* is the time scale.
* is the average time scale.
* is the scaled Schoenfeld residuals for covariate k.
* is the information matrix elements for covariate k.
* d is the event times.

The test statistics that exceed 5% critical values are considered as the proportional hazard assumptions are not violated.

Schoenfeld residual test examines whether the residuals for each covariate show an ordered association with time, which would indicate a violation of the assumption. In the context of a study on head and neck cancer, Schoenfeld residuals were used to evaluate the PH assumption for the covariates included in the Cox proportional hazards model. The analysis involved both global and individual Schoenfeld tests. The global test evaluates the total commitment of all covariates to the PH assumption, while individual tests assess each covariate separately. In this study, if the global and individual tests yielded p-values greater than 0.05, the conventional threshold for statistical significance, results convey no significant evidence to reject the null hypothesis that the PH assumption holds, suggesting the covariates align with the probability threshold.

Graphical evaluation of Schoenfeld residuals provided further validation. The residuals were plotted against time, with each red dot representing a residual value for a specific covariate at a particular time point. The red dots appeared randomly scattered around zero, showing no clear trend or regular variation. Additionally, the smoothed solid line, which represents the overall trend of residuals, stayed nearly horizontal and close to zero which indicates no strong relationship between residuals and time. The dashed lines represent the 95% confidence intervals, and enclose the majority of the residuals without displaying a clear pattern of deviation from proportionality. A non-zero slope in the residual plot would signify time dependence for a covariate, implying its effect on the hazard changes over time, thus violating the PH assumption.

##### **3.7.4 Selection of Variables Based on Akaike Information Criterion (AIC)**

In Cox proportional hazards regression, the selection of variables based on the Akaike Information Criterion (AIC) is a widely used method for model optimization. AIC helps in balancing model fit and complexity by penalizing the involvement of redundant variables. Variable selection based on the AIC can be performed using forward selection, backward elimination, or stepwise selection.

Forward selection begins with a model that contains no covariates. Variables are added sequentially, starting with the covariate that most improves the model fit, specifically, the one that results in the greatest reduction in the AIC. After each addition, the AIC is recalculated. The process continues by adding the variable that improves the model until no further improvements are possible. The goal is to include only variables that significantly explain the variance in the dependent variable while avoiding overfitting by excluding irrelevant covariates. Backward elimination starts with a model that includes all candidate covariates. In this method, the least significant variable identified as the one with the highest p-value or the smallest contribution to explaining the variability in the dependent variable is removed first. After each removal, the AIC is recalculated. This process continues until no variable can be removed without increasing the AIC. Backward elimination aims to simplify the model by removing variables that add little value while preserving a good fit. Stepwise selection combines both forward and backward approaches, iterating between adding and removing covariates depending on their effect on the AIC. The AIC is calculated by,

Where,

* ln(L) is the log-likelihood of the model
* k is the number of parameters (covariates) in the model.

These methods identify the most significant variables while minimizing overfitting and maintaining model parsimony. In this study, the backward elimination method is used to identify only the variables that significantly explain the outcome while avoiding overfitting by excluding irrelevant variables.

#### **3.8 Lasso-Based Cox Regression Model**

Lasso-based Cox regression is an extension of the Cox proportional hazard (Cox PH) model. This approach is especially valuable in high-dimensional datasets where a large number of covariates may be irrelevant or redundant. The regularization in Lasso-based Cox regression is achieved through the use of the L1-norm, which penalizes the absolute values of the regression coefficients. It helps to prevent overfitting and improves the model's interpretability by retaining only the most important covariates. The L1-norm penalty is the sum of the absolute values of the coefficients. i.e,

Where βi​ are the coefficients of the covariates Xi.

This penalty forces some coefficients to become exactly zero, effectively removing the corresponding variables from the model. The Lasso-based Cox model modifies the standard Cox PH likelihood function as,

* is the partial log-likelihood function of the Cox Proportional Hazards model.
* λ is the regularization parameter that controls the amount of penalty applied to the regression coefficients. A larger λ results in stronger regularization, leading to more coefficients being forced to zero.

##### **3.8.1 Cross-Validation**

Cross-validation is a technique used to assess a model's performance and select optimal parameters, such as the regularization parameter λ in Lasso-based Cox regression. One of the most typical forms of cross-validation is k-fold cross-validation. In this method, the whole dataset is separate into k equally sized folds or subsets. The procedure begins by allocating one-fold to serve as the test set, while the remaining k−1 folds are used to train the model. This training and testing process is repeated k times, with each fold being used once as the test set. By changing the test set through all k folds, we ensure that every data point is used for both training and testing across the iterations. To evaluate the model's performance, a specific performance metric is calculated for each fold. Depending on the nature of the task, this metric could be the concordance index (C-index), commonly used for survival models, or mean squared error (MSE), frequently used in regression tasks. After obtaining the performance metrics from all k iterations, we average these results to generate a comprehensive estimate of the model's performance. The cross-validation error is calculated as the average error across all k folds, and the value of λ that minimizes this error is selected. The general formula for cross-validation error for a given model is,

Where,

* is the regularization parameter.
* k is the number of folds.
* is the error (or performance measure) calculated for the model trained on all folds except fold i and evaluated on fold i.

In the context of Lasso-based Cox regression, the role of cross-validation is particularly important in determining the optimal value of the regularization parameter λ. The process involves testing multiple candidate values for λ and selecting the one that leads to the lowest prediction error or the highest performance metric. The cross-validation error is computed as the average error over all k folds, and the value of λ that minimizes this error becomes the chosen parameter. Cross-validation helps ensure that a model works well with new data and prevents overfitting. It does this by choosing the best model parameters.

#### **3.9 Methods used for evaluating the model performance**

When assessing predictive models for survival in patients with head and neck cancer, it is essential to evaluate their ability to discriminate among patients based on their results, such as survival rates. To conduct a detailed analysis of this aspect, we applied two key statistical metrics: the C-index and the time-dependent area under the curve (AUC). Both the C-index and the time-dependent AUC play complementary roles in assessing the efficiency of Cox proportional hazards (PH) models and Lasso-based Cox regression models. Together, these metrics provide a more comprehensive evaluation of model performance in survival analysis.

##### **3.9.1 Concordance Index**

The C-index, or Concordance Index, is a metric used to evaluate the discriminatory power of survival models, such as the Cox Proportional Hazards (Cox PH) model or Lasso-Cox regression. It measures the model's predictive accuracy by assessing how often the predictions match the actual outcomes. Specifically, the C-index evaluates how well the model can differentiate between individuals based on their predicted risk of an event over time. The C-index works by comparing pairs of individuals. For each pair, the model assigns a predicted risk score, which indicates the likelihood of the event occurring for each individual. The actual event times for both individuals are also recorded. The value of the C-index ranges from 0.5 to 1. A C-index value of 0.5 suggests that the model has no discriminatory ability, while a value of 1.0 indicates perfect discrimination.

##### **3.9.2 Time-Dependent AUC**

The time-dependent Area Under the Curve (AUC) is a significant metric for determining the effectiveness of survival models, such as the Cox Proportional Hazards (Cox PH) and Lasso-Cox regression models. While the concordance index (C-index) measures overall discriminatory ability, the time-dependent AUC provides a specific evaluation of model performance at various time points. This is especially important in survival analysis, where the likelihood of events can change over time.

The time-dependent AUC is calculated at different points during the follow-up period. It compares the predicted risk scores against the actual survival outcomes, indicating the model’s effectiveness in identifying individuals who will experience an event at each specific time. This metric is derived from the receiver operating characteristic (ROC) curve, which plots the true positive rate (sensitivity) against the false positive rate (specificity) at different thresholds. Unlike traditional AUC, the time-dependent AUC considers the time aspect, making it particularly useful for survival analysis. By examining the time-dependent AUC at multiple durations, we can better understand how well the model distinguishes between individuals who have experienced the event by that certain time and those who have not. This dynamic evaluation helps recognize periods when the model is most effective, providing a more thorough assessment of its predictive accuracy over time.

## Chapter 4

## Materials

### **4.1 Study Design**

Retrospective cohort study

### **4.2 Study Setting**

The study has been initiated by the Department of Cancer Registry and Epidemiology of Malabar Cancer Centre, Thalassery.

### **4.3 Study Variables**

In this study, we collected and analyzed a comprehensive set of variables from patients diagnosed with Head and Neck cancer. The variables include demographic, clinical, pathological treatment, and follow-up information, crucial for understanding the factors influencing patient outcomes. Below is a detailed description of each variable included in this study.

* Age: Age of the patient recorded in years at the time of diagnosis.
* Sex: The biological classification of the patient as male or female.
* Date of diagnosis: The date when the head and neck cancer was diagnosed.
* Primary site: The organ where the cancer first develops. Common sites in head and neck cancer include the hypopharynx, larynx, mouth, nasopharynx, oropharynx, pharynx, tongue, and tonsils.
* Histology: The tumour types in head and neck cancer include SCC (Squamous Cell Carcinoma), carcinoma, adenocarcinoma, and others.
* Grade: The tumour grade indicates the aggressiveness of cancer cells. It may be well-differentiated, moderately differentiated, poorly differentiated, or undifferentiated.
* Tumour size: The dimension of the tumour categorized as 1, 1A, 1B, 2, 3, 4, 4A, 4B, and 4C.
* Nodal Status: Indicates whether cancer has spread to lymph nodes or not.
* Metastasis Status: Indicates whether cancer has spread to other parts of the body.
* Composite Stage: The combined assessment of cancer growth which is classified into stages 1 to 4.
* Treatment given prior to registration at RI: Indicates whether the patient received any treatment before registration at the research institute.
* Type of treatment given: The treatment received by the patient including surgery (S), chemotherapy (C), radiotherapy (R), hormone therapy (H), S+C, S+R, R+C, and S+R+C.
* Tobacco Chewing: Indicate the habit of chewing tobacco.
* Smoking Habit: Indicate the habit of smoking.
* Alcohol Consumption: Indicate the consumption of alcohol.
* Date of last follow-up: The last date on which the patient’s status was recorded.
* Date of Death: The date of the patient’s death if applicable.
* Status: Current status of the patient. 1 if the patient dies and 0 if the patient is alive or lost to follow-up.
* Lifetime in month: The duration from the date of diagnosis to the date of death if the patient has expired, or to the date of last follow-up, measured in months.

### **4.4 Inclusion and Exclusion Criteria**

1162 Head and Neck cancer patients who registered according to Hospital Based Cancer Registry at Malabar Cancer Centre during 2017-2019

4 patients were excluded as they were duplicate entries

186 patients were excluded as they didn't receive cancer directed treatment at MCC

976 patients were included

972 Head and Neck cancer patients were included in the study

## Chapter 5

## Results

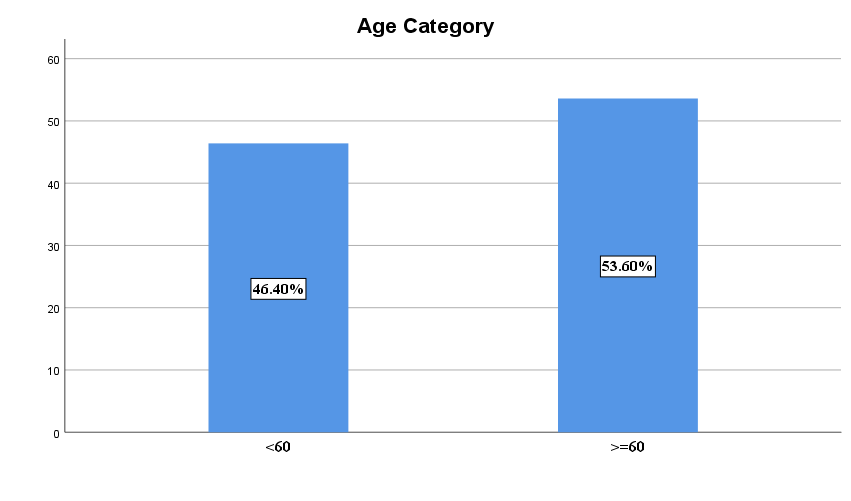
**5.1 Baseline Characteristics of Head and Neck Cancer Patients**

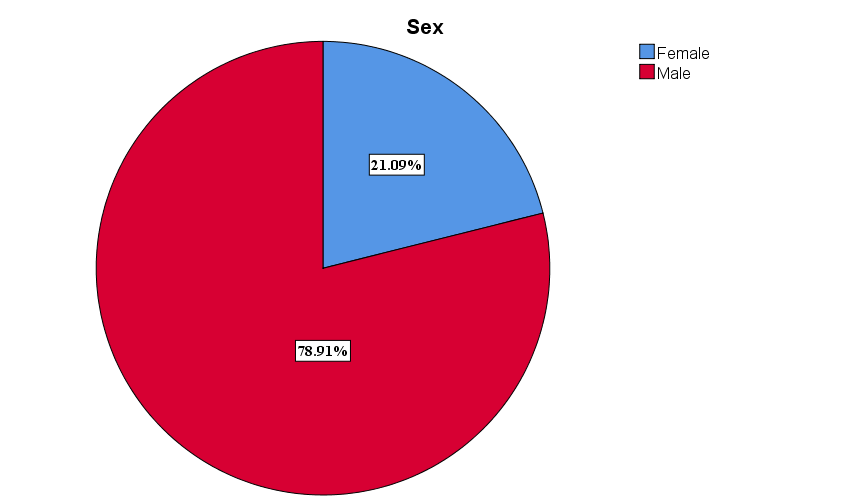
The study involved 972 patients with Head and Neck Cancer patients who received cancer-directed treatment at Malabar Cancer Centre, Thalassery. Among these approximately 40.43% (393) of patients died due to Head and Neck Cancer and other causes. The remaining 59.57% (579) of patients were right censored. The mean and median age at diagnosis were 58.84 and 60 respectively. The median follow-up time is 72 months with 95% confidence interval (70.294, 73.706). The below table presents the characteristics of patients diagnosed with head and neck cancer.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Category** | **Frequency** | **Percentage** |
| **Age Category** | <60 | 451 | 46.4 |
| >=60 | 521 | 53.6 |
| **Sex** | Female | 205 | 21.09 |
| Male | 767 | 78.91 |
| **Marital Status** | Divorced | 6 | 0.62 |
| Married | 897 | 92.28 |
| Unmarried | 17 | 1.75 |
| Widowed | 52 | 5.35 |
| **Religion** | Christian | 101 | 10.39 |
| Hindu | 673 | 69.24 |
| Jain | 1 | 0.1 |
| Muslim | 197 | 20.27 |
| **Education** | College and above | 39 | 4.01 |
| Illiterate | 97 | 9.98 |
| Literate | 162 | 16.67 |
| Middle | 206 | 21.19 |
| Primary | 283 | 29.12 |
| Secondary | 171 | 17.59 |
| Others | 1 | 0.1 |
| Unknown | 13 | 1.34 |
| **Primary site** | Hypopharynx | 68 | 7 |
| Larynx | 199 | 20.47 |
| Mouth | 332 | 34.16 |
| Nasopharynx | 9 | 0.93 |
| Oropharynx | 85 | 8.74 |
| Pharynx | 4 | 0.41 |
| Tongue | 263 | 27.06 |
| Tonsil | 12 | 1.23 |
| **Histology** | Adeno Carcinoma | 8 | 0.82 |
| Carcinoma | 8 | 0.82 |
| Others | 16 | 1.65 |
| SCC | 933 | 95.99 |
| Unknown | 7 | 0.72 |
| **Grade** | MD | 355 | 36.52 |
| PD | 50 | 5.14 |
| WD | 314 | 32.3 |
| Undifferentiated | 2 | 0.21 |
| Unknown | 251 | 25.82 |
| **Staging** | TNM | 971 | 99.9 |
| Others (specify) | 1 | 0.1 |
| **Tumour Size** | 1 | 170 | 17.49 |
| 1A | 19 | 1.95 |
| 1B | 11 | 1.13 |
| 2 | 325 | 33.44 |
| 3 | 211 | 21.71 |
| 4 | 47 | 4.84 |
| 4A | 157 | 16.15 |
| 4B | 20 | 2.06 |
| 4C | 1 | 0.1 |
| Unknown | 11 | 1.13 |
| **Nodal Status** | Yes | 467 | 48.05 |
| No | 496 | 51.03 |
| Unknown | 9 | 0.93 |
| **Metastasis Status** | Yes | 8 | 0.82 |
| No | 909 | 93.52 |
| Unknown | 55 | 5.66 |
| **Composite Stage** | Stage 1 | 175 | 18 |
| Stage 2 | 204 | 20.99 |
| Stage 3 | 365 | 37.55 |
| Stage 4 | 224 | 23.05 |
| Not applicable | 1 | 0.1 |
| Unknown | 3 | 0.31 |
| **Treatment given prior to**  **registration at RI** | Yes | 99 | 10.19 |
| No | 872 | 89.71 |
| Unknown | 1 | 0.1 |
| **Cancer direct treatment at RI** | Yes | 948 | 97.53 |
| Incomplete Treatment | 24 | 2.47 |
| **Treatment Given** | Surgery (S) | 247 | 25.41 |
| Chemotherapy (C) | 5 | 0.51 |
| Radiotherapy (R) | 165 | 16.98 |
| Hormone Therapy(H) | 2 | 0.21 |
| S+C | 6 | 0.62 |
| S+R | 214 | 22.02 |
| R+C | 193 | 19.86 |
| S+R+C | 140 | 14.4 |
| **Tobacco Chewing** | Yes | 393 | 40.43 |
| No | 534 | 54.94 |
| Unknown | 45 | 4.63 |
| **Smoking Habit** | Yes | 549 | 56.48 |
| No | 390 | 40.12 |
| Unknown | 33 | 3.4 |
| **Alcohol Consumption** | Yes | 506 | 52.06 |
| No | 433 | 44.55 |
| Unknown | 33 | 3.4 |
| **Status** | Alive | 579 | 59.57 |
| Death | 393 | 40.43 |

5.1 Baseline characteristics of patients with head and neck cancer

The table comprehensively analyses the patient characteristics including demographics, clinical, pathological, treatment, and follow-up information. The age distribution shows that the slight majority of patients are aged 60. 521 patients (53.6%) were above the median age, and 451 patients (46.4%) were below the median age. Majority of the patients are male (78.9%) and females are only 21.09% of the population under study.

5.1 Graphical Display of Age Category



5.2 Graphical Display of Sex

The primary site is categorized into hypopharynx (7%), larynx (20.47%), mouth (34.16%), nasopharynx (0.93%), oropharynx (8.74%), pharynx (0.41%), tongue (27.06%), and tonsil (1.23%). Histology shows that Squamous Cell Carcinoma is most common in patients (95.99%) and remaining adenocarcinoma (0.82%), carcinoma (0.82%), others (1.65%), and a small percentage of unknown (0.72%). Tumour grades are distributed as, well-differentiated (32.3%), moderately differentiated (36.52%), poorly differentiated (5.14%), and undifferentiated (0.21%), with a significant portion of unknown (25.82%). Nodal status shows 48.05% positive and 51.03% negative, with a small portion unknown (0.93%). Metastasis is absent in more patients (93.52%) with a small portion (0.82%) having metastasis, and unknown (5.66%). Composite staging shows the distribution as stage 1 (18%), stage 2 (20.99%), stage 3 (37.55%), stage 4 (23.05%), and a small portion of not applicable (0.1%), and unknown (0.31%). Only 10.19% took treatment before registration at MCC and the remaining 89.71 didn’t take any treatment for head and neck cancer. Most patients (97.53%) completed cancer directed treatment given at MCC with a small percentage of incomplete treatments (2.47%). Patients received various treatment combinations such as surgery alone (25.41%), radiotherapy alone (16.98%), chemotherapy alone (0.51%), hormone therapy alone (0.21%), surgery + radiotherapy (22.02%), surgery + chemotherapy (0.62%), radiotherapy + chemotherapy (19.86%) and surgery + radiotherapy + chemotherapy (14.4%). 40.43% of patients reported chewing tobacco, 54.94% did not, and 4.63% were categorized as unknown. Regarding smoking habits, 56.48% of patients were smokers, 40.12% were non-smokers, and 3.4% fell into the unknown category. 52.06% consume alcohol while 44.55% do not and 3.4% the unknown.

**5.2 Survival Rate of Head and Neck Cancer Patients Using Kaplan-Meier Method**

Kaplan-Meier method of survival analysis of head and neck cancer patients gives significant insights into survival probabilities over time. In this study, the Kaplan Meier estimates suggest an overall 6-year survival rate of 57.5%. Also, Survival rates at 1, 3, and 5-year intervals are estimated to be 88.1%, 70%, and 61.5%, respectively. The dataset’s median follow-up time obtained through the reverse Kaplan-Meier method is 72 months with a 95% confidence interval of 70.29 to 73.71 months. The table elaborates on these rates of each year with corresponding lower and upper 95% confidence intervals for precision. The Kaplan-Meier curve depicts the changing pattern of survival rates highlighting the layered aspects of survival experience.

|  |  |  |  |
| --- | --- | --- | --- |
| **Time (Year)** | **Survival Rate** | **Lower limit** | **Upper limit** |
| 1 | 88.1 | 86 | 90.1 |
| 3 | 70 | 67.1 | 73 |
| 5 | 61.5 | 58.4 | 64.8 |
| 6 | 57.5 | 54.3 | 60.9 |

5.2 Survival rate of Head and Neck cancer patients by Kaplan Meier method

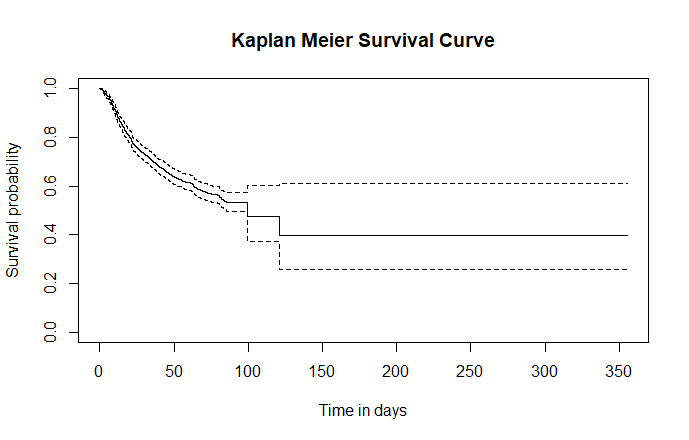


Figure 5.3 Survival curve of head and neck cancer patients using Kaplan Meier method

**5.3 Comparison of** **Survival Rate of Head and Neck Cancer Patients using the Log Rank Test**

The Log Rank Test evaluates the survival probabilities among Head and Neck Cancer patients while considering the covariates age category, sex, primary site, histology, tumour grade, tumour size, nodal status, metastasis status, composite stage, treatment given, tobacco chewing, smoking habit, and alcohol consumption. For each covariate, a p-value is calculated to determine its statistical significance which helps to identify factors that significantly influence survival outcomes in this patient population. Based on the Log Rank Test results, the covariates sex and Histology are not statistically significant, as their p-values exceed the 0.05 threshold. All other variables, including age category, primary site, grade, tumour size, nodal status, metastasis status, composite stage, treatment given, tobacco chewing, smoking habit, and alcohol consumption, are statistically significant factors influencing Head and Neck Cancer patients.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Category** | **Frequency** | **P value** |
| **Age Category** | <60 | 451 | 0.002\* |
| >=60 | 521 |
| **Sex** | Female | 205 | 0.159 |
| Male | 767 |
| **Primary site** | Hypopharynx | 68 | 0.0001\* |
| Larynx | 199 |
| Mouth | 332 |
| Nasopharynx | 9 |
| Oropharynx | 85 |
| Pharynx | 4 |
| Tongue | 263 |
| Tonsil | 12 |
| **Histology** | Adeno Carcinoma | 8 | 0.253 |
| Carcinoma | 8 |
| Others | 16 |
| SCC | 933 |
| Unknown | 7 |
| **Grade** | MD | 355 | 0.0003\* |
| PD | 50 |
| WD | 314 |
| Undifferentiated | 2 |
| Unknown | 251 |
| **Tumour Size** | 1 | 171 | 0.002 x 10-11 \* |
| 1A | 19 |
| 1B | 11 |
| 2 | 324 |
| 3 | 211 |
| 4 | 47 |
| 4A | 157 |
| 4B | 20 |
| 4C | 1 |
| Unknown | 11 |
| **Nodal Status** | Yes | 467 | 0.002 x 10-10 \* |
| No | 496 |
| Unknown | 9 |
| **Metastasis Status** | Yes | 8 | 0.01\* |
| No | 909 |
| Unknown | 55 |
| **Composite Stage** | Stage 1 | 176 | 0.003 x 10-13 \* |
| Stage 2 | 203 |
| Stage 3 | 365 |
| Stage 4 | 224 |
| Not applicable | 1 |
| Unknown | 3 |
| **Treatment Given** | Surgery (S) | 246 | 0.005 x 10-10 \* |
| Chemotherapy (C) | 5 |
| Radiotherapy (R) | 166 |
| Hormone Therapy(H) | 2 |
| S+C | 6 |
| S+R | 214 |
| R+C | 193 |
| S+R+C | 140 |
| **Tobacco Chewing** | Yes | 393 | 0.006 x 10-4 \* |
| No | 534 |
| Unknown | 45 |
| **Smoking Habit** | Yes | 549 | 0.009 x 10-5 \* |
| No | 390 |
| Unknown | 33 |
| **Alcohol Consumption** | Yes | 506 | 0.002 x 10-4 \* |
| No | 433 |
| Unknown | 33 |

5.3 Survival Rate of Head and Neck Cancer Patients using the Log Rank Test

**5.4** **Univariate Cox Proportional Hazard Model for Head and Neck Cancer Survival Data**

The univariate Cox proportional hazard model analysis below table gives detailed insights into the factors influencing survival among head and neck cancer patients. The effect is quantified through the estimate, p-value, hazard ratio, and confidence interval (CI).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Covariate** | **Category** | **Estimate** | **P value** | **Hazard ratio** | **95% CI** | |
| **Lower CI** | **Upper CI** |
| **Age Category** | **<60** | **Reference** | | | | |
| >=60 | 0.235 | 0.0215 \* | 1.265 | 1.035 | 1.546 |
| **Sex** | **Female** | **Reference** | | | | |
| Male | 0.181 | 0.160 | 1.198 | 0.931 | 1.541 |
| **Primary Site** | **Hypopharynx** | **Reference** | | | | |
| Larynx | -0.361 | 0.086 | 0.697 | 0.462 | 1.052 |
| Mouth | -0.205 | 0.292 | 0.815 | 0.557 | 1.193 |
| Nasopharynx | -1.621 | 0.110 | 0.198 | 0.027 | 1.446 |
| Oropharynx | 0.369 | 0.096 | 1.447 | 0.936 | 2.235 |
| Pharynx | 0.510 | 0.397 | 1.666 | 0.511 | 5.433 |
| Tongue | -0.408 | 0.0453 \* | 0.665 | 0.344 | 0.992 |
| Tonsil | -0.197 | 0.657 | 0.821 | 0.344 | 1.961 |
| **Histology** | **Adeno Carcinoma** | **Reference** | | | | |
| Carcinoma | -0.605 | 0.508 | 0.546 | 0.091 | 3.279 |
| SCC | 0.118 | 0.839 | 1.125 | 0.361 | 3.505 |
| Unknown | 1.038 | 0.174 | 2.824 | 0.631 | 12.636 |
| Others | -0.228 | 0.755 | 0.796 | 0.190 | 3.335 |
| **Grade** | **MD** | **Reference** | | | | |
| PD | 0.448 | 0.0247 \* | 1.566 | 1.059 | 2.315 |
| WD | -0.410 | 0.00124 \* | 0.663 | 0.517 | 0.851 |
| Undifferentiated | 0.067 | 0.947 | 1.069 | 0.150 | 7.639 |
| Unknown | -0.122 | 0.337 | 0.885 | 0.690 | 1.136 |
| **Tumour Size** | **1** | **Reference** | | | | |
| 1A | 0.055 | 0.917 | 1.057 | 0.376 | 20.969 |
| 1B | -0.198 | 0.783 | 0.821 | 0.198 | 3.405 |
| 2 | 0.666 | 0.0003 \* | 1.947 | 1.347 | 2.814 |
| 3 | 1.039 | 0.005x 10-5 \* | 2.826 | 1.942 | 4.114 |
| 4 | 1.039 | 0.009 x 10-2 \* | 2.825 | 1.678 | 4.755 |
| 4A | 1.293 | 0.005 x 10-8 \* | 3.644 | 2.477 | 5.360 |
| 4B | 1.214 | 0.0004 \* | 3.367 | 1.716 | 6.604 |
| 4C | 3.638 | 0.0003 \* | 38.033 | 5.137 | 281.601 |
| Unknown | -0.264 | 0.717 | 0.768 | 0.184 | 3.199 |
| **Nodal Status** | Yes | 0.750 | 0.005 x 10-10 \* | 2.116 | 1.725 | 2.595 |
| **No** | **Reference** | | | | |
| Unknown | -1.307 | 0.194 | 0.271 | 0.038 | 1.942 |
| **Metastasis Status** | Yes | 1.269 | 0.0049 \* | 3.557 | 1.470 | 8.611 |
| **No** | **Reference** | | | | |
| Unknown | 0.085 | 0.689 | 1.088 | 0.719 | 1.648 |
| **Composite Stage** | **Stage 1** | **Reference** | | | | |
| Stage 2 | 0.543 | 0.0096 \* | 1.722 | 1.141 | 2.597 |
| Stage 3 | 0.985 | 0.001 x 10-4 \* | 2.677 | 1.858 | 3.855 |
| Stage 4 | 1.437 | 0.005 x 10-11 \* | 4.209 | 2.893 | 6.123 |
| Not applicable | -14.330 | 0.995 | 0.005 x 10-4 | 0 | Inf |
| Unknown | -14.310 | 0.992 | 0.006 x 10-4 | 0 | Inf |
| **Treatment given prior to registration at RI** | Yes | -0.246 | 0.163 | 0.002 x 10-3 | 0.553 | 1.105 |
| **No** | **Reference** | | | | |
| Unknown | -0.13 | 0.988 | 0.781 | 0 | Inf |
| **Treatment Given** | Surgery (S) | -1.949 | 0.001 \* | 0.142 | 0.144 | 0.458 |
| **Chemotherapy (C)** | **Reference** | | | | |
| Radiotherapy (R) | -1.153 | 0.051 | 0.316 | 0.099 | 1.007 |
| Hormone Therapy(H) | -0.630 | 0.586 | 0.533 | 0.055 | 5.124 |
| S+C | -0.312 | 0.683 | 0.732 | 0.164 | 3.274 |
| S+R | -0.874 | 0.136 | 0.418 | 0.132 | 1.317 |
| R+C | -0.709 | 0.226 | 0.492 | 0.156 | 1.552 |
| S+R+C | -0.707 | 0.231 | 0.493 | 0.155 | 1.567 |
| **Tobacco Chewing** | Yes | 0.124 | 0.243 | 1.132 | 0.920 | 1.393 |
| **No** | **Reference** | | | | |
| Unknown | 0.997 | 0.002 x 10-4 \* | 2.709 | 1.854 | 3.958 |
| **Smoking Habit** | Yes | 0.215 | 0.0454 \* | 1.240 | 1.004 | 1.531 |
| **No** | **Reference** | | | | |
| Unknown | 1.187 | 0.005 x 10-5 \* | 3.278 | 2.137 | 5.027 |
| **Alcohol Consumption** | Yes | 0.143 | 0.175 | 1.154 | 0.938 | 1.418 |
| **No** | **Reference** | | | | |
| Unknown | 1.137 | 0.001x 10-4 \* | 3.117 | 2.041 | 4.761 |

5.4 Results of Univariate Cox Model

Using univariate Cox analysis, researchers found that patients aged 60 and older have a much higher risk of death compared to those under 60, with a 26.5% increase in risk. Gender does not significantly affect the risk of death, although males have a 19.8% higher risk than females. Among tumour locations, patients with tongue tumours have a significantly lower risk of death compared to those with tumours in the hypopharynx. When looking at different tumour types compared to Adeno Carcinoma, there are differences in death risk, but they aren't statistically significant. Carcinoma is linked to a 45.4% lower risk, while Squamous Cell Carcinoma (SCC) shows a 12.5% higher risk. Other tumour types have a 20.4% lower risk, but these differences aren’t statistically significant. Poorly differentiated tumours carry a higher risk of death compared to moderately differentiated tumours, while well-differentiated tumours are linked to a lower risk. Larger tumours, especially sizes 4A, 4B, and 4C, significantly increase the risk of death.

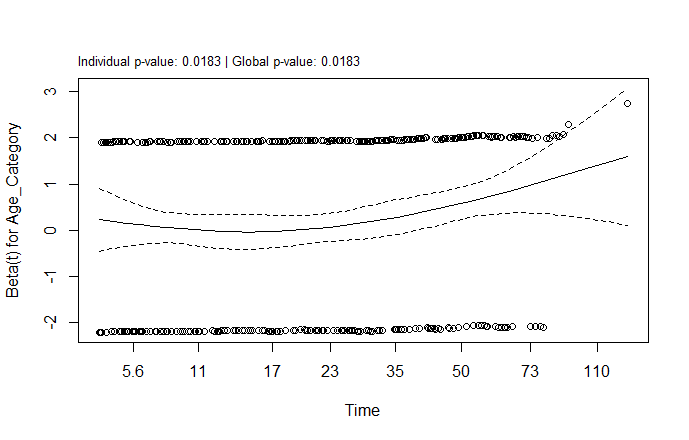
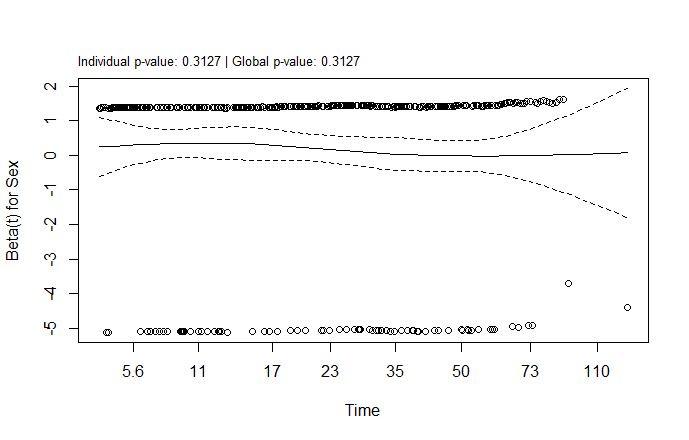
Having cancer spread to lymph nodes and other parts of the body also raises the risk of death. Stages of cancer, particularly Stage 4, show a much higher risk of death compared to Stage 1. Patients who had surgery have a significantly lower risk of death, while other treatment methods do not show notable differences.

For those who chew tobacco, the risk of death is similar to non-chewers, even though it is 13.2% higher, but not statistically significant. On the other hand, smokers have a 24% higher risk of death compared to non-smokers, which is statistically significant. Alcohol consumption does not significantly impact death risk, as drinkers show a 15.4% higher risk than non-drinkers, but this difference isn't statistically significant either.

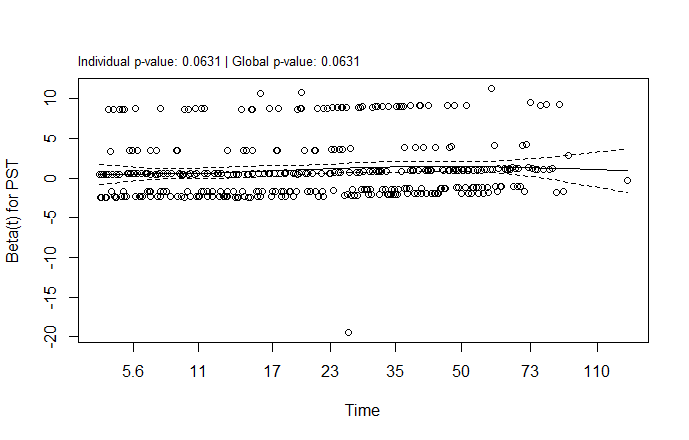
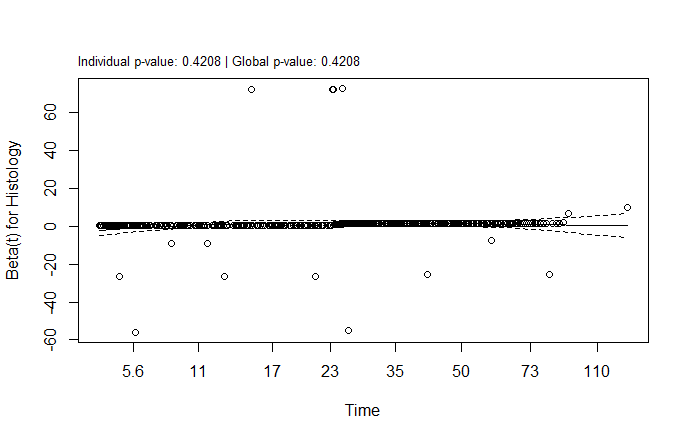
Using univariate Cox analysis, Age, Primary Site (PST), Tumour Grade, Tumour Size, Nodal Status, Metastasis Status, Composite Stage, Treatment given, and Smoking Habit are the significant factors affecting survival.

**Proportional Hazard Assumption for Head and Neck Cancer Survival Data**

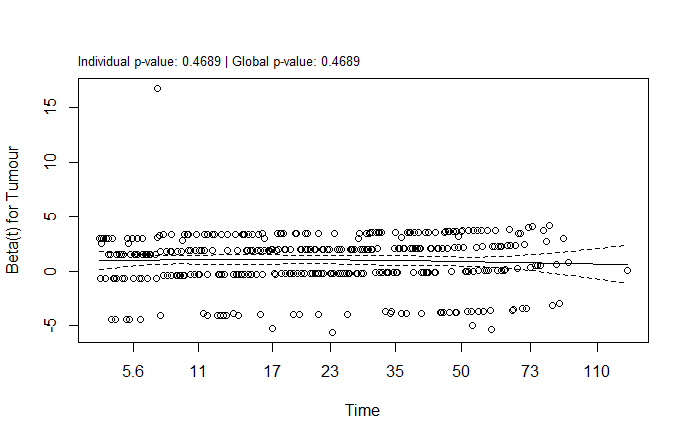
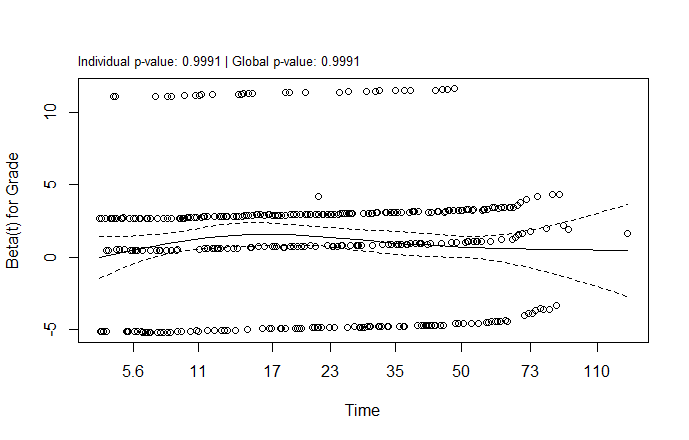
Based on the Schoenfeld residual plots and corresponding statistical test results, the proportional hazard assumption holds for most of the covariates in the head and neck cancer dataset. However, the covariate age category violated the proportional hazard assumption. The deviations observed are indicated in the Schoenfeld residual plot given below.



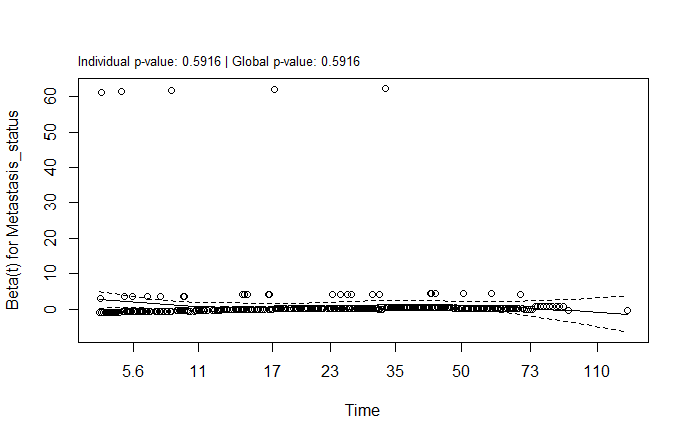
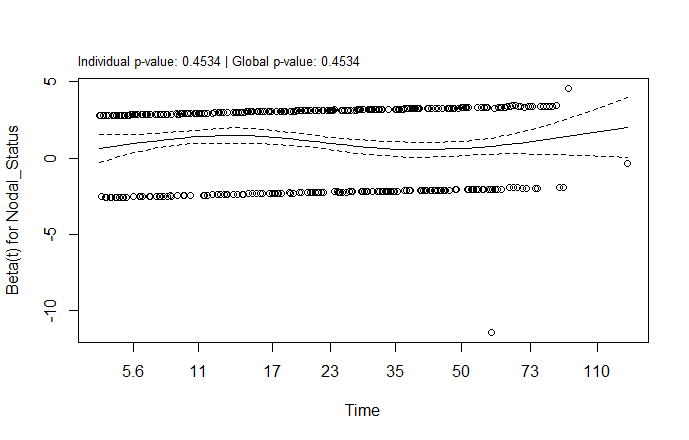
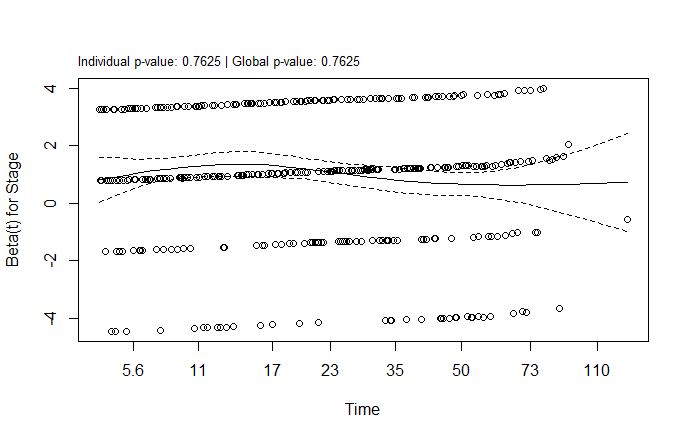
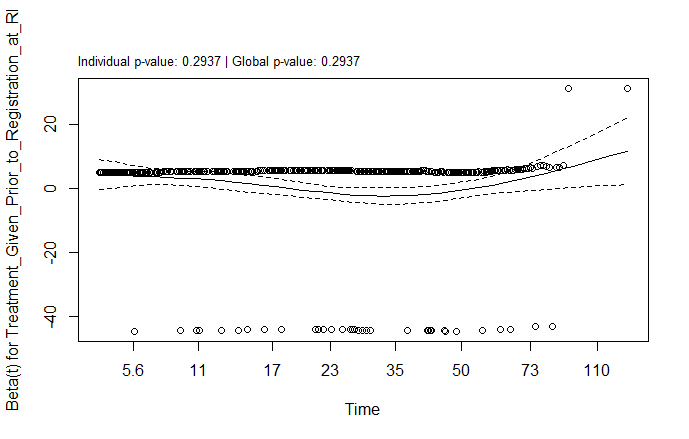
5.4 PH assumption plot for Age category 5.5 PH assumption plot for Sex

****

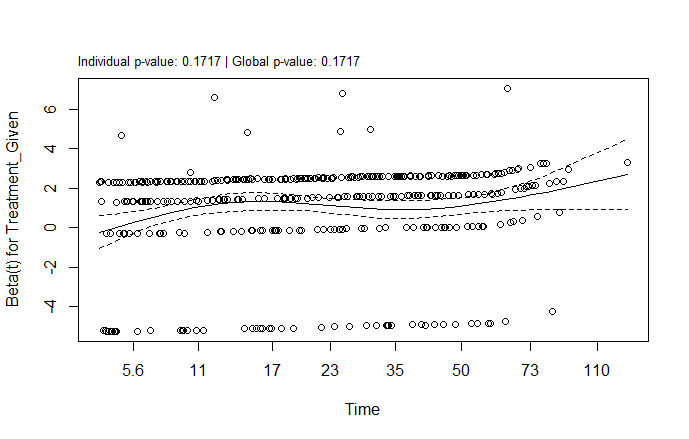
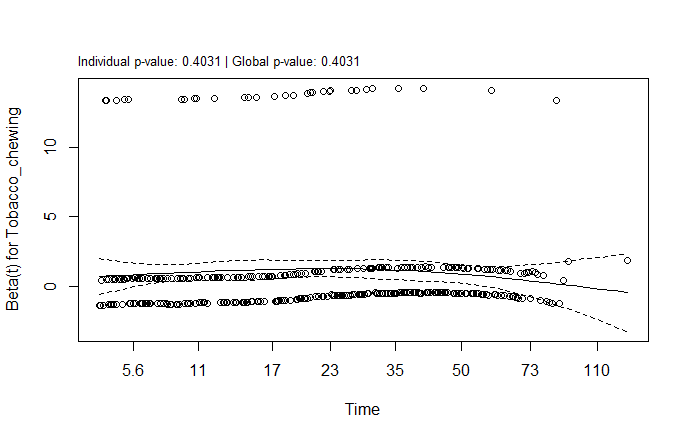
5.6 PH assumption plot for primary site 5.7 PH assumption plot for Histology

****

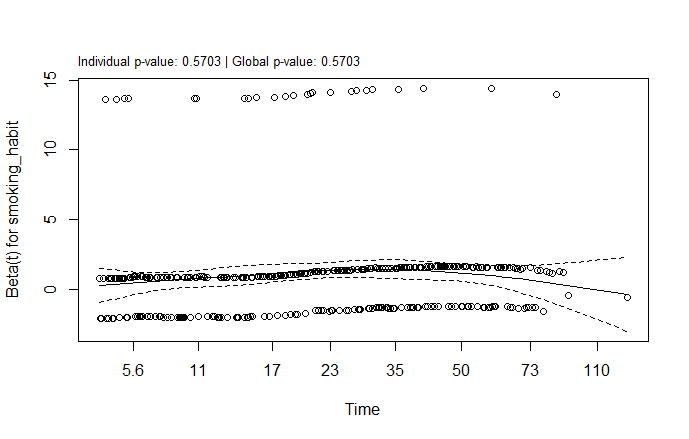
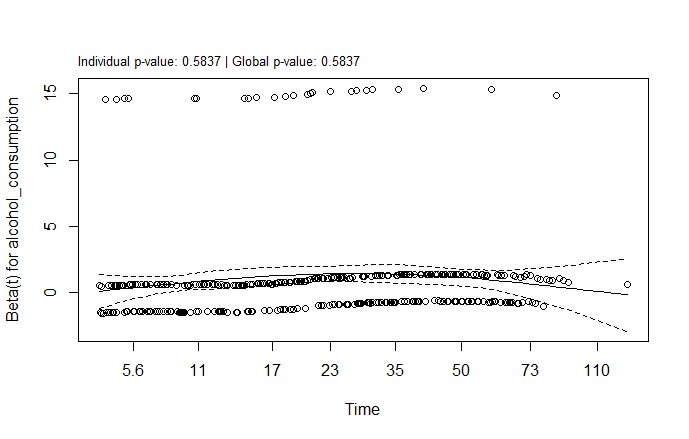
5.8 PH assumption plot for Tumour grade 5.9 PH assumption for Tumour size

**** 5.10 PH assumption for Nodal Status 5.11 PH assumption for Metastasis Status

5.12 PH assumption plot for Composite stage 5.13 PH assumption plot for treatment given before registration at RI



5.14 PH assumption plot for type of treatment 5.15 PH assumption plot for tobacco given chewing

5.16 PH assumption plot for Smoking habit 5.17 PH assumption plot for Alcohol Consumption

**5.6** **Selection of Variables for Multivariate Cox Proportional Hazard Model**

|  |  |
| --- | --- |
| **Model** | **AIC** |
| Sex + Primary site + Histology + Grade + Tumour +Nodal Status +Metastasis status+ Stage+ Treatment Given Prior to Registration at RI+ Treatment Given +Tobacco chewing + smoking habit+ alcohol consumption | 5002.7 |
| Sex + Primary site + Grade + Tumour +Nodal Status +Metastasis status+ Stage+ Treatment Given Prior to Registration at RI+ Treatment Given +Tobacco chewing + smoking habit+ alcohol consumption | 4996.79 |
| Sex + Primary site + Grade + Tumour +Nodal Status +Metastasis status+ Treatment Given Prior to Registration at RI+ Treatment Given +Tobacco chewing + smoking habit+ alcohol consumption | 4992.68 |
| Sex + Primary site + Grade + Tumour +Nodal Status +Metastasis status+ Treatment Given +Tobacco chewing + smoking habit+ alcohol consumption | 4990.22 |
| Sex + Primary site + Grade + Tumour +Nodal Status +Metastasis status+ Treatment Given + smoking habit+ alcohol consumption | 4988.05 |
| Sex + Primary site + Grade + Tumour +Nodal Status +Metastasis status+ Treatment Given + smoking habit | 4986.19 |
| Primary site + Grade + Tumour +Nodal Status +Metastasis status+ Treatment Given + smoking habit | 4984.54 |
| Primary site + Grade + Tumour +Nodal Status + Treatment Given + smoking habit | 4983.64 |

5.5 Selection of Variables to Conduct Multivariate Cox PH model based on AIC

Using the backward elimination method, based on the Akaike Information Criterion (AIC)values calculated for various models in the above table, we can determine the suitable model by comparing these values. The model with all variables, Age category, Sex, primary site, histology, tumour grade, tumour size, nodal status, metastasis status, composite stage, treatment given prior to registration at RI, type of treatment given, tobacco chewing, smoking habit, and alcohol consumption has an AIC of 5002.7. Further simplifying the model by removing histology results in a model with AIC decreases to 4996.79. Similarly, by removing the stage, treatment given prior to registration at RI, tobacco chewing, sex, alcohol consumption, and metastasis status in each step AIC reduces to 4992.68, 4990.22, 4988.05, 4986.19, and 4984.54 respectively. The most simplified model includes only the primary site, tumour grade, tumour size, nodal status, treatment given, and smoking habit, which achieves the lowest AIC of 4983.64.

**5.7** **Multivariate Cox Proportional Hazard Model for Head and Neck Cancer Survival Data**

The multivariate Cox proportional hazard model provides several significant factors influencing the outcome of head and neck cancer patients. The effect is quantified through the estimate value, p-value, hazard ratio, and 95% confidence interval.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Covariate** | **Category** | **Estimate** | **P value** | **Hazard ratio** | **95% CI** | |
| **Lower CI** | **Upper CI** |
| **Primary Site** | **Hypopharynx** | **Reference** | | | | |
| Larynx | -0.01 | 0.964 | 0.99 | 0.638 | 1.536 |
| Mouth | -0.058 | 0.797 | 0.943 | 0.605 | 1.471 |
| Nasopharynx | -1.793 | 0.081 | 0.167 | 0.022 | 1.249 |
| Oropharynx | 0.372 | 0.102 | 1.45 | 0.929 | 2.265 |
| Pharynx | 1.386 | 0.032\* | 3.998 | 1.126 | 14.188 |
| Tongue | 0.053 | 0.824 | 1.054 | 0.662 | 1.68 |
| Tonsil | -0.033 | 0.941 | 0.967 | 0.401 | 2.335 |
| **Grade** | **MD** | **Reference** | | | | |
| PD | 0.442 | 0.031\* | 1.557 | 1.041 | 2.327 |
| WD | -0.321 | 0.015\* | 0.726 | 0.56 | 0.94 |
| Undifferentiated | 0.16 | 0.875 | 1.174 | 0.16 | 8.594 |
| Unknown | 0.004 | 0.974 | 1.004 | 0.774 | 1.303 |
| **Tumour Size** | **1** | **Reference** | | | | |
| 1A | -0.354 | 0.516 | 0.702 | 0.242 | 2.038 |
| 1B | -0.52 | 0.479 | 0.594 | 0.141 | 2.513 |
| 2 | 0.333 | 0.089 | 1.396 | 0.949 | 2.052 |
| 3 | 0.538 | 0.01\* | 1.713 | 1.135 | 2.585 |
| 4 | 0.574 | 0.048\* | 1.735 | 1.006 | 3.131 |
| 4A | 0.848 | 0.0001\* | 2.334 | 1.504 | 3.623 |
| 4B | 0.753 | 0.042\* | 2.122 | 1.027 | 4.387 |
| 4C | 3.466 | 0.0009\* | 32.012 | 4.122 | 248.61 |
| Unknown | -0.165 | 0.852 | 0.848 | 0.149 | 4.82 |
| **Nodal Status** | Yes | 0.421 | 0.007\* | 1.523 | 1.192 | 1.945 |
| **No** | **Reference** | | | | |
| Unknown | -1.547 | 0.221 | 0.213 | 0.018 | 2.535 |
| **Treatment Given** | Surgery (S) | -1.645 | 0.007\* | 0.193 | 0.058 | 0.644 |
| **Chemotherapy (C)** | **Reference** | | | | |
| Radiotherapy (R) | -0.947 | 0.123 | 0.388 | 0.117 | 1.29 |
| Hormone Therapy(H) | -1.75 | 0.142 | 0.174 | 0.017 | 1.793 |
| S+C | -0.63 | 0.419 | 0.532 | 0.115 | 2.458 |
| S+R | -1.001 | 0.096 | 0.367 | 0.113 | 1.193 |
| R+C | -1.034 | 0.09 | 0.355 | 0.107 | 1.175 |
| S+R+C | -1.081 | 0.075 | 0.339 | 0.103 | 1.113 |
| **Smoking Habit** | Yes | 0.065 | 0.588 | 1.067 | 0.843 | 1.351 |
| **No** | **Reference** | | | | |
| Unknown | 1.267 | 0.005 x 10^-5\* | 3.55 | 2.251 | 5.6 |

5.6 Results of Multivariate Cox PH Model

The multivariate Cox proportional hazard model shows that among various tumour types, the pharynx has 4 times more risk of death compared to those with the hypopharynx. Among tumour grades, well-differentiated tumours have a 27.4% lower risk of death compared to moderately differentiated ones while poorly differentiated tumours have a 55.7% higher risk. Well-differentiated tumours are associated with better outcomes compared to moderately differentiated ones, while poorly differentiated ones are riskier than moderately differentiated ones. When comparing tumour sizes 3,4,4A,4B, and 4C with tumour size 1 as a reference, it significantly results in a higher risk of death. Especially 4C has a high value of hazard ratio (32.012) among them.

Having cancer spread to lymph nodes has a 52.3% higher risk of death compared to those with don’t spread to lymph nodes. Patients who had surgery have a significantly lower risk of death at a rate of 80.7% compared to those who had chemotherapy. In the smoking category, the unknowns have a significantly higher risk of death than the patients who don’t smoke.

Using multivariate Cox proportional analysis, the covariates primary site, tumour grade, tumour size, nodal status, type of treatment given, and smoking habit have a statistically significant impact on survival of head and neck cancer patient’s survival.

**5.8** **Cross Validation and Selection of Optimal Regularisation Parameter for Lasso-based Cox Regression Model**

By applying cross-validation, the optimal value for the regularisation parameter (λ) obtained for the full model is 0.03024 which minimizes the cross-validation error. This optimal value gives the best balance between the model fit and complexity.

Below is a given plot of cross-validation for the Lasso-based Cox model that visualizes the optimal value of the regularisation parameter. The plot is nothing other than log λ Vs partial likelihood deviance. Initially, the graph starts with a higher value of partial likelihood of deviance. Gradually, the value of the partial likelihood of deviance decreases to a certain point; after that, the curve flattens or increases. The optimal value of the regularisation parameter is chosen where the partial likelihood of deviance reaches its minimum.

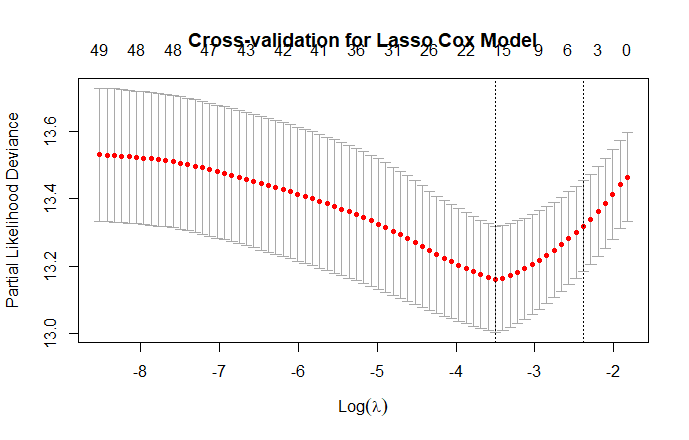


Figure 5.18 Cross Validation plot for Lasso Cox model

**5.9** **Selection of Variables for Lasso-Based Cox Regression Model**

The chosen optimal regularisation parameter value is used to select variables and to avoid overfitting. This approach effectively selects the covariates that affect survival and avoids irrelevant variables. The indication of relevance is based on coefficient value. A non-zero coefficient indicates that the corresponding variable influences the patient’s outcome. Others are removed from the final model. Also, the direction of the coefficient is important, since it indicates increasing hazard if the coefficient is positive and vice versa.

|  |  |  |
| --- | --- | --- |
| **Covariates** | **Levels** | **Coefficient** |
| **Age category** | <60 | -0.163 |
| **>=60** | **Reference** |
| **Primary Site** | **Hypopharynx** | **Reference** |
| Larynx | - |
| Mouth | - |
| Nasopharynx | -0.404 |
| Oropharynx | -0.248 |
| Pharynx | - |
| Tongue | - |
| Tonsil | - |
| **Grade** | **Moderately differentiated** | **Reference** |
| Poorly differentiated | 0.17 |
| Well differentiated | -0.133 |
| Undifferentiated | - |
| Unknown | - |
| **Tumour Size** | **1** | **Reference** |
| 1A | - |
| 1B | - |
| 2 | - |
| 3 | 0.052 |
| 4 | - |
| 4A | 0.298 |
| 4B | - |
| 4C | - |
| Unknown | - |
| **Nodal Status** | Yes | 0.321 |
| **No** | **Reference** |
| Unknown | -0.099 |
| **Metastasis Status** | Yes | 0.256 |
| **No** | **Reference** |
| Unknown | - |
| **Stage** | **Stage 1** | **Reference** |
| Stage 2 | - |
| Stage 3 | - |
| Stage 4 | 0.248 |
| Not applicable | - |
| Unknown | - |
| **Treatment given** | Surgery (S) | -0.567 |
| **Chemotherapy (C)** | **Reference** |
| Radiotherapy (R) | - |
| Hormone therapy (H) | - |
| S+C | - |
| S+R | - |
| R+C | - |
| S+R+C | - |
| **Tobacco Chewing** | Yes | - |
| **No** | **Reference** |
| Unknown | 0.094 |
| **Smoking habit** | Yes | - |
| **No** | **Reference** |
| Unknown | 0.719 |
| **Alcohol consumption** | Yes | - |
| **No** | **Reference** |
| Unknown | 0.008 x 10-12 |

5.7 Selection of Variables for Lasso-Cox Model by Cross Validation

The coefficient of those under 60 is -0.163, which indicates that those under 60 have a lower hazard compared to those aged 60 and above. When analyzing the covariate sex with females as the reference, the non-zero coefficient indicates that sex is irrelevant. Nasopharyngeal and oropharyngeal cancer has relevance when compared to hypopharyngeal cancer where the nasopharynx shows a lower risk while the oropharynx shows a higher risk. Histological categories do not show significant contributions to the result. Among tumour grades, well-differentiated show a higher risk while poorly differentiated show a lower risk when compared to moderately differentiated. Tumour sizes 3 and 4A show a positive relevance indicting a higher risk than tumour size 1. The presence of nodal involvement and metastasis presents a higher risk when compared to those without such involvement. Using Stage 1 as the reference, Stage 4 contributes a significant effect on survival and indicates a higher risk. Among the various treatments given to the patients, only those with surgery (S) show a lower risk of death. Regarding lifestyle factors such as tobacco chewing, smoking, and alcohol consumption, the unknowns have an increasing risk of death when compared to non-users.

The variables age category, primary site, tumour grade, tumour size, nodal status, metastasis status, composite stage, type of treatment given, tobacco chewing, smoking habit, and alcohol consumption significantly contribute to the model’s prediction. Other variables, sex, histology, and treatment given before registration at RI are irrelevant as they do not contribute to the model’s predictive accuracy.

**5.10** **Performance of Models for Head and Neck Cancer Survival Data**

|  |  |  |
| --- | --- | --- |
| **Performance metrics** | **Cox PH** | **Lasso -Based Cox** |
|
| **C-Index** | 0.676 | 0.673 |
| **AUC** | 0.720 | 0.717 |

The study uses the C index and time-dependent AUC to evaluate the performance of two survival models: the Cox proportional hazard model and the Lasso-based Cox regression model. The result does not indicate a markable difference between both models. The C index value ranges from 0.5 to 1. A higher value of the C index and AUC indicates a better model performance.

5.8 Performance of Cox PH model and Lasso Cox model

The result shows that the C-index of the Cox model is 0.676 and the AUC is 0.72 indicating the model has a good ability to distinguish between different levels of risk in a dataset. Similarly, the C-index and AUC of the Lasso-based Cox model are 0.673 and 0.717 respectively which are very close to that of the Cox model. Both the C-index and AUC suggest minimal differences between the models. Either the Cox model or the Lasso-based Cox model can be considered for suitable prediction for this dataset.

## Chapter 6

## Discussion and Conclusion

In this study, we compared the performance of the Cox proportional hazard model and the Lasso-based Cox regression models for patients diagnosed with head and neck cancer. We found that each model achieves better discriminative ability.

The study involved 972 patients with Head and Neck Cancer treated at Malabar Cancer Centre, Thalassery. Of these, 40.43% died from cancer or other causes, while 59.57%were right censored. The mean age at diagnosis was 58.84 and the median follow-up time was 72 months. The dataset includes variables such as age, sex, tumour site, histology, tumour grade, composite stage, tumour size, nodal status, metastasis status, type of treatment given, and factors like smoking, alcohol consumption, and tobacco chewing.

In a study on head and neck cancer in India by Sonali Bagal et. Al (2023) the factors influencing the survival of head and neck cancer patients include sex and lifestyle factors like tobacco, smoking, and alcohol consumption [9]. In our study, the covariates sex, tobacco chewing, and alcohol consumption are not statistically significant in both the Cox PH model and the Lasso-based Cox regression model. In the study of epidemiology and survival analysis of head and neck cancer by Amit Badola et. Al (2023) was conducted on 574 patients from 2016 to 2020, resulting in a median follow-up of 21 months. The 18-month survival probability in that study is 58.5% and we got 57.5% in 72 months. Badola also found that age and gender are not significant, but tobacco is found to be significant since that study included 84.8% of the population with tobacco chewers [29]. In most studies, the Lasso-based Cox regression model is better than that of the Cox PH model. We found that each model has a better predictive accuracy. Therefore, either the Cox model or the lasso-based Cox model can be considered for risk prediction.

In this study, Kaplan Meier provides survival probability of 88.1, 70, 61.5, and 54.3 over 1, 3, 5, and 6 years respectively. Log Rank test suggests Age category, Primary Site, Grade, Tumour size, Nodal Status, Metastasis Status, Composite stage, Treatment given, Tobacco Chewing, Smoking Habit, and Alcohol Consumption as the statistically significant factors influencing survival of head and neck cancer patients.

When we go through the proportional hazard assumption for the Cox model only the age category significantly violates the assumptions. Univariate Cox results show that covariates such as sex, histology, treatment given before registration at RI, tobacco chewing, and alcohol consumption did not significantly affect survival. The model selected for multivariate Cox PH using the stepwise backward elimination method includes primary site, tumour grade, tumour size, nodal status, treatment given, and smoking habit. Also, these variables show statistical significance in multivariate analysis.

The lasso-based Cox model suggests prognostic factors influencing the outcome of head and neck cancer patients by regularising the covariates in the dataset using a penalty. The optimal regularisation parameter obtained is 0.03024. The model after regularisation has prognostic factors such as age category, primary site, tumour grade, tumour size, nodal status, metastasis status, composite stage, type of treatment given, tobacco chewing, smoking habit, and alcohol consumption. Sex, histology, and treatment given before registration at RI do not significantly affect the survival of head and neck cancer patients.

When comparing these two models, the Cox PH and the lasso-based Cox model, we found that both the C-index and AUC values are similar for each model. Cox has a C-index of 0.676 whereas the lasso model has 0.673. When comparing AUC, Cox has an AUC of 0.72 whereas the Lasso Cox model has 0.717 which also indicates each model has a better predictive accuracy.

### **6.1 Conclusion**

The study aims to identify the prognostic factors influencing the survival of head and neck cancer patients treated at Malabar Cancer Centre from 2017 to 2019 using the Cox PH model and the Lasso-based Cox model along with the comparison of these two models. The median follow-up obtained is 72 months with a survival probability of 54.3% using the Kaplan-Meier method of survival analysis. Using the Cox PH model the factors influencing survival are primary site, tumour grade, tumour size, nodal status, type of treatment given, and smoking habit while the Lasso-based Cox model included age category, metastasis status, composite stage, tobacco chewing, and alcohol consumption. Each model showcases higher accuracy in prediction with a higher C-index and AUC. Despite previous studies suggesting that the Lasso-based Cox model performs better in prediction. This study indicates that both the Cox model and the Lasso-based Cox model can be used for survival prediction for this dataset.

**References**

1. American Cancer Society- Head and neck cancer retrieved from <https://www.cancer.org/cancer/types/head-neck-cancer.html>, accessed on December 2024.
2. National Cancer Institute- Head and neck cancer-patient version retrieved from https://www.cancer.gov/types/head-and-neck <https://www.cancer.gov/types/head-and-neck>, accessed on December 2024
3. Macmillan cancer support- Head and neck cancer retrieved from <https://www.macmillan.org.uk/cancer-information-and-support/head-and-neck-cancer>, accessed on December 2024
4. Blot, W.J., et al. "Tobacco and alcohol use about head and neck cancers." Journal of the National Cancer Institute, vol. 87, no. 12, 1995, pp. 936-942.
5. John Hopkins Medicine- Head and neck cancer retrieved from <https://www.hopkinsmedicine.org/health/conditions-and-diseases/head-and-neck-cancer>, accessed on December 2024.
6. [Mayo Clinic](https://www.mayoclinic.org/diseases-conditions/head-and-neck-cancers/symptoms-causes/syc-20360799)​ [Memorial Sloan Kettering Cancer Center](https://www.mskcc.org/cancer-care/types/head-neck/risk-prevention-screening) – head and neck cancers retrieved from https://www.mayoclinic.org/diseases-conditions/head-and-neck-cancers/symptoms-causes/syc-20354171, accessed on December 2024
7. Freddy Bray et al., “Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.” Journal of American Cancer Society, Vol.74(3)
8. Malabar Cancer Centre, Hospital-Based Cancer Registry (HBCR) Report 2022 retrieved from <https://www.mcc.kerala.gov.in/publications/mcc-pgiosr-hbcr-annual-report-2022>, accessed on December 2024
9. Bagal S, et al. Head and neck cancer burden in India: an analysis from published data of 37 population-based cancer registries. A journal on Ecancermedicalscience 2023, Vol. 17
10. Jemal A, et al. Global Cancer Statistics, 2020. A Cancer Journal of the American Cancer Society. 2021;71(3).
11. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic concepts and first analyses. Br J Cancer. 89, 232–238 (2003).
12. Etikan I, Abubakar S, Alkassim R. The Kaplan-Meier estimate in survival analysis. Biom Biostat Int J. 2017;5(2):55-59
13. Sullivan, L. (2016). Survival analysis. Boston University, School of Public Health.
14. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1996 Jan;58(1):267-88.
15. Zhou, D., Liu, X., Wang, X., et al. (2021). A prognostic nomogram based on LASSO Cox regression in patients with alpha-fetoprotein-negative hepatocellular carcinoma following non-surgical therapy. BMC Cancer, 21, 246.
16. Xu, Y., et al. (2022). A LASSO-based survival prediction model for patients with synchronous colorectal carcinomas based on SEER. Translational Cancer Research, 11(8), 2795.
17. Bai, L., & Gillen, D. (2017). Survival analysis via Cox proportional hazards additive models. Encyclopedia with Semantic Computing and Robotic Intelligence, 1(01), 1650003.
18. Sun, S., et al. (2020). Development and validation of an immune‐related prognostic signature in lung adenocarcinoma. European Journal of Cancer Care, 9(16), 5960-5975.
19. Wang, Q., et al. (2022). Nomogram established on account of Lasso-Cox regression for predicting recurrence in patients with early-stage hepatocellular carcinoma. Frontiers in Immunology, 13, 1019638.
20. Zhou, J., et al. (2024). Suicide among lymphoma patients. Journal of Affective Disorders, 360, 97-107.
21. Jia, X., et al. (2024). Identification of AURKA as a biomarker associated with cuproptosis and ferroptosis in HNSCC. International Journal of Molecular Sciences, 25(8), 4372.
22. Sui, S., et al. (2020). An immune cell infiltration-based immune score model predicts prognosis and chemotherapy effects in breast cancer. Theranostics, 10(26), 11938.
23. Shahraki, H. R., Salehi, A., & Zare, N. (2015). Survival prognostic factors of male breast cancer in Southern Iran: a LASSO-Cox regression approach. Asian Pacific Journal of Cancer Prevention, 16(15), 6773-6777.
24. Hahn, G., et al. (2024). Polygenic hazard score models for the prediction of Alzheimer's free survival using the Lasso for Cox's proportional hazards model. Genetic Epidemiology, 49(1).
25. Cai, X., et al. (2020). Development and validation of a novel model for predicting the 5‐year risk of type 2 diabetes in patients with hypertension: A retrospective cohort study. BioMed Research International, 2020, 9108216.
26. Wang, Y., et al. (2021). Prognostic implications of immune-related eight-gene signature in pediatric brain tumours. Brazilian Journal of Medical and Biological Research, 54(7), e10612.
27. Arends, C. R., et al. (2020). Optimizing survival predictions of hypopharynx cancer: development of a clinical prediction model. The Laryngoscope, 130(9), 2166-2172.
28. Li, S., et al. (2021). Determination of a six-gene prognostic model for cervical cancer based on WGCNA combined with LASSO and Cox-PH analysis. World Journal of Surgical Oncology, 19, 277.
29. Badola, A., et al. (2023). Epidemiology and survival analysis of head and neck cancer: results from a comprehensive care center in North India. Oral Oncology Reports, 6, 100022.

**APPENDIX**

library(survival)

library(survminer)

library(glmnet)

data=lung

head(lung)

km\_fit = survfit(Surv(time,status) ~ 1,data=data)

summary(km\_fit)

plot(km\_fit,xlab = "Time in days",ylab = "Survival probability",main="Kaplan Meier Survival Curve")

km\_sex = survfit(Surv(time,status) ~ sex,data = data)

print(summary(km\_sex,times = 60))

plot(km\_sex,col = c("blue","red"),lwd=2,xlab = "Time in days",ylab = "Survival probability",main="Kaplan Meier Survival Curves by Sex")

legend("topright", legend=c("Male","Female"),col = c("blue","red"),lwd=2)

surv\_object = Surv(time = data$time, event = data$status)

log\_rank\_test = survdiff(surv\_object ~ sex, data = data)

#print(log\_rank\_test)

chi\_sq<- log\_rank\_test$chisq

df<- length(log\_rank\_test$n) - 1

p\_value<- pchisq(chi\_sq, df = df, lower.tail = FALSE)

cat("Chi-squared:", chi\_sq, "P-value:", p\_value, "\n")

cox\_model = coxph(Surv(time,status) ~ age + sex + ph.ecog+ph.karno + pat.karno + meal.cal + wt.loss,data = data)

summary(cox\_model)

cox\_test=cox.zph(cox\_model)

cox\_test

ggcoxzph(cox\_test)

library(MASS)

model = coxph(Surv(time,status)~ inst+sex + ph.ecog+ph.karno + pat.karno + meal.cal + wt.loss,data=data\_clean)

stepAIC(model,direction="both",trace=FALSE)

# Create a model matrix and survival object

X <- model.matrix(~ age + sex + ph.ecog + ph.karno + pat.karno + meal.cal + wt.loss - 1, data = data\_clean)

y <- Surv(data\_clean$time, data\_clean$status)

# Fit Lasso Cox model with glmnet

fit = glmnet(X, y, family = "cox")

# Plot the regularization path

plot(fit)

lasso\_cv<- cv.glmnet(X, y, family = "cox", alpha = 1)

cat("Optimal lambda:", lasso\_cv$lambda.min, "\n")

# Extract coefficients at optimal lambda

lasso\_coefs<- coef(lasso\_cv, s = lasso\_cv$lambda.min)

print(lasso\_coefs)

plot(lasso\_cv, main = "Cross-validation for Lasso Cox Model")

# Make predictions for the risk scores using the Lasso model

lasso\_risk\_scores<- predict(lasso\_cv, newx = X, s = lasso\_cv$lambda.min, type = "link")

# Fit a Cox model using the predicted risk scores as a covariate

lasso\_cox\_model<- coxph(Surv(data\_clean$time, data\_clean$status) ~ lasso\_risk\_scores)

summary(lasso\_cox\_model)

# C-index for the Cox model

cox\_cindex<- summary(cox\_model)$concordance[1]

cat("C-index for Cox model:", cox\_cindex, "\n")

# AIC for the Cox model

cox\_aic<- AIC(cox\_model)

cat("AIC for Cox model:", cox\_aic, "\n")

# Predict risk scores from the Lasso-Cox model

lasso\_risk\_scores<- predict(lasso\_cv, newx = X, s = lasso\_cv$lambda.min, type = "link")

# C-index for Lasso Cox model

c\_index<- survConcordance(Surv(data\_clean$time, data\_clean$status) ~ lasso\_risk\_scores)

lasso\_cindex<- c\_index$concordance

cat("C-index for Lasso Cox model:", lasso\_cindex, "\n")

log\_lik<- logLik(lasso\_cox\_model)

# Calculate AIC

k <- length(lasso\_coefs[lasso\_coefs != 0])  # Number of non-zero coefficients

lasso\_aic<- -2 \* log\_lik + 2 \* k

cat("AIC for Lasso Cox model:", lasso\_aic, "\n")