Lung Cancer

Batoul Blaybel

February 7, 2025

Contents

1 Introduction

Lung cancer is the leading cause of cancer-related death worldwide, with non-small cell lung cancer (NSCLC) accounting for approximately 85% of all lung cancer cases. Its incidence has significantly increased in many countries, particularly in the Western world, over the past few decades. Smoking remains the primary cause of lung cancer, with tobacco use responsible for the vast majority of cases. However, other risk factors, such as exposure to environmental pollutants, radon, and occupational carcinogens, also contribute to lung cancer development, with varying prevalence across different regions. While smoking rates have decreased in some countries, the burden of lung cancer remains high, especially among former and current smokers. The risk of lung cancer is directly related to the intensity and duration of smoking, with heavy smokers (those consuming more than 20 cigarettes per day for 20 or more years) having a significantly higher risk of developing the disease. In countries like the United States and France, lung cancer accounts for tens of thousands of deaths annually. Recent studies suggest that tumor stage at diagnosis may be influenced by smoking history, with lung cancer in smokers often being diagnosed at a later stage compared to non-smokers, which negatively impacts prognosis. This cohort is a prospective observational study that collected data on the clinical features and treatment allocations of patients with newly diagnosed lung cancer. In this cohort, smoking-related lung cancer was predominant. We leveraged this large prospective cohort to compare the clinical features at diagnosis, treatment approaches, and outcomes between smoking-related and non-smoking-related lung cancer.

2 Materials and methods

2.1 Handling Missing Data

Missing values were present in the dataset for several variables. The nature of the missingness was analyzed using a Generalized Linear Model (GLM), which determined that missingness in variables such as Criteria, Abstinence Status,Difuse Cancer, and Age Group was Missing Not At Random (MNAR), while missingness in other variables was Missing At Random (MAR).

- **Imputation for MAR Variables**

  Imputation for MAR Variables was performed using the Multiple Imputation by Chained Equations (MICE) method:

  - Quantitative Variables: Imputed using the Predictive Mean Matching (PMM) method.
  - Qualitative Variables with Two Levels: Imputed using logistic regression (logreg method).
  - Qualitative Variables with More Than Two Levels: Imputed using polynomial regression (polyreg method).[3]

- **Imputation for MNAR Variables**

  For MNAR variables, specific imputation methods were applied:

  - A Bayesian logistic regression model was fitted using the `stan_glm` function. The model incorporated predictors such as survival in days, HBV, HCV, cancer stages, and treatment variables, etc.
  - Predicted probabilities for the missing variable were calculated.
  - Missing values were imputed using a binomial sampling approach based on these probabilities. [2]

- **Imputation for Age**

  For missing values in Age, a Random Forest model was applied using the Age Group variable as a predictor.

2.2 Data Processing and Descriptive Statistics

The dataset consisted of 894 patients, including both categorical and numerical variables. Descriptive statistics were calculated to summarize the dataset.

- **Categorical Variables:** Frequency distributions and proportions were calculated. Additionally, correlations between categorical variables were assessed using Cramér's V, which measures the strength of association between pairs of categorical variables. The results revealed no high correlation between the variables, indicating that multicollinearity among the categorical variables was not a concern.

- **Numerical Variables:** Means and standard deviations were calculated , along with 95% confidence intervals for the means.For specific quantitative variables such as TumorSize and SurvivalInDays, histograms and boxplots were generated to examine their distributions.

The Freedman-Diaconis rule was used to determine optimal bin widths for histogram plots.

2.3 Statistical Modeling and Risk Analysis

- **Relative Risk (RR):** Relative risks were calculated for categorical variables, with Group A used as the reference category for multi-level variables.RR quantifies the likelihood of death in one group compared to another.

- **Logistic Regression:** A logistic regression model assessed the association between `TumorSize` and `DeathStatus`.

- **Linear Regression:** A linear regression model evaluated the relationship between `TumorSize` and `SurvivalInDays`.

2.4 Lead Time Bias Estimation

Lead time is the period between the moment a disease becomes detectable and when it becomes clinically manifest. Lead time bias was assessed using two standard methodologies:

- **Schwartz Method:** The Schwartz method is a statistical approach designed to correct for lead time bias in cancer screening studies. It models the preclinical detectable phase of cancer, assuming tumor progression follows an exponential growth pattern. By simulating tumor growth over time, this method estimates the lead time gained through earlier detection by screening and adjusts survival estimates to prevent overestimation of survival benefits.

  - **Doubling Time (DT):** Simulated using a log-normal distribution. If the simulated DT exceeds 440 days, it is replaced with the mean value.
  - **Tumor Size of Screened Patients (DS):** Simulated using a log-normal distribution. If the simulated DS value exceeds 100 mm or is less than 10 mm, it is replaced with the mean value.
  - **Tumor Size of Unscreened Patients (DNS):** Simulated similarly to DS. If the simulated DNS value exceeds 200 mm or is less than 10 mm, it is replaced with the mean value.

  The lead time (LT_{sim}) was calculated as:

$$LT_{\text{sim}} = 3 \times DT_{\text{sim}} \times \frac{\log(DNS_{\text{sim}}/DS_{\text{sim}})}{\log(2)}$$

  Negative lead times were replaced with the mean to ensure interpretability. Simulations were repeated 1000 times to generate robust estimates, and summary statistics were derived from the resulting lead time distributions.

- **Duffy Method:**

  The Duffy method estimates lead time bias by simulating the period between early detection by screening and clinical diagnosis. Lead time is calculated using survival days (`SurvivalInDays`), a specified time threshold (`t`), and the mean sojourn time. Exponential functions model the relationship between these variables to adjust survival estimates and eliminate bias caused by early detection.

  - **Sojourn Time:** Simulated using a triangular distribution with a minimum value of 10, a mode value of 12, and a maximum value of 14.to represent the duration during which the disease is detectable by screening before clinical diagnosis occurs [1]
  - **Lead Time (LT):** The adjustment applied to observed survival time, calculated based on survival days and the mean sojourn time.

  The corrected survival time (CorrectedSurvival12) is calculated as:

$$\text{CorrectedSurvival12} = \text{SurvivalInDays} - \text{LeadTimeM22}$$

  where LeadTimeM22 is the lead time adjustment based on the Duffy method. The corrected survival time helps eliminate bias introduced by early detection in survival analysis.

2.5 Survival Analysis

To evaluate factors affecting corrected survival, the following analysis were conducted:

- **Kaplan-Meier Method:** Survival curves were estimated for different patient groups using the Kaplan-Meier estimator. and differences between groups were evaluated using the log-rank test.
- **Cox Proportional Hazards Model:** A multivariable Cox regression model was applied to assess the effect of covariates on corrected survival times. The model is defined as:

$$h(t|X) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) \tag{1}$$

where $h(t|X)$ is the hazard function at time t given covariates X, $h_0(t)$ is the baseline hazard, and β_i are the coefficients representing the effect of covariate X_i on the hazard. Hazard ratios ($\exp(\beta_i)$) were calculated to quantify the effect sizes of covariates.

To identify the best-fitting model, a backward selection approach was then applied.

3 Results and Discussion

Variable	Levels	Percentage (%)	RR	Mean ± SD
Gender	Male	82.6	1.14	–
	Female	17.4	–	–
Age	–	–	–	67.35 ± 11.24
AgeGroup	$\leq$ 70 years	56.6	0.93	–
	$>$ 70 years	43.4	–	–
Smoker	Yes	65.1	1.18	–
	No	34.9	–	–
AbstinenceStatus	Abstinent	55.1	0.96	–
	Non-Abstinent	44.9	–	–
HBV	Yes	9.3	0.90	–
	No	90.7	–	–
HCV	Yes	12.5	0.80	–
	No	87.5	–	–
Other	Yes	8.3	0.96	–
	No	91.7	–	–
Screening	Screened	22.6	0.83	–
	Unscreened	77.4	–	–
TumorSize	–	–	–	61.34 ± 61.05
Diabetes	Yes	33.1	0.95	–
	No	66.9	–	–

Table 1: Summary of Variables (Part 1)

Variable	Levels	Percentage (%)	RR	Mean ± SD
AlcoholConsumption	Yes	35.3	0.99	–
	No	64.7	–	–
Thrombosis	Yes	28.6	1.28	–
	No	71.4	–	–
Criteria	In	28.5	0.65	–
	Out	71.5	–	–
CancerStages	A	15.8	1.00	–
	B	5.6	1.34	–
	C	47.3	1.32	–
	D	31.3	1.53	–
DiffuseCancer	Yes	15.9	1.27	–
	No	84.1	–	–
MetastaticCancer	Yes	12.8	1.17	–
	No	87.2	–	–
CurativeTreatment	With curative intent	20.5	0.48	–
	Without curative intent	79.5	–	–
Treatment1	Yes	10.2	1.14	–
	No	89.8	–	–
Treatment2	Yes	8.2	1.03	–
	No	91.8	–	–
Treatment3	Yes	16.4	0.92	–
	No	83.6	–	–
Treatment4	Yes	17.2	0.98	–
	No	82.8	–	–
Treatment5	Yes	37.4	0.95	–
	No	62.6	–	–
Treatment6	Yes	7.9	1.07	–
	No	92.1	–	–
DeathStatus	Death	74.0	–	–
	No	26.0	–	–
SurvivalInDays	–	–	–	331.76 ± 387.58

Table 2: Summary of Variables (Part 2)

Demographic Characteristics

The demographic and clinical characteristics of lung cancer patients in this study are summarized in Tables 1 and 2. The cohort consisted predominantly of males (82.6%), with a mean age of 67.35 ± 11.24 years. Approximately 56.6% of the patients were aged 70 years or younger, and 65.1% were current smokers. Among those assessed for smoking abstinence, 55.14% reported being abstinent.

Clinical Factors

Regarding clinical factors, 22.59% of patients were screened by ultrasound, whereas the majority (77.4%) were unscreened. The mean tumor size was 61.34 ± 61.05 mm, with 71.5% of patients having a tumor size greater than 50 mm and 28.5% having a tumor size less than 50 mm. Diabetes was present in 33.1% of the cohort, while alcohol consumption was reported by 35.3%. Thrombosis was identified in 28.6% of patients. Viral infections were also noted, with 9.3% of patients testing positive for HBV and 12.5% for HCV. Difuse cancer was present in 15.9% of cases, and metastatic cancer was identified in

12.8%.

Cancer Stages

Cancer stages showed a predominance of advanced disease, with 47.3% of cases classified as advanced stages (Stage C), while 31.3% were in terminal stages (Stage D). Early stages (Stage A and B) accounted for 21.4% of patients.

Treatment Data

Treatment data revealed that most patients (37.4%) received Treatment 5, followed by 17.2% receiving Treatment 4, and 16.4% receiving Treatment 3. A significant proportion (79.5%) received treatment without curative intent, while only 20.5% underwent treatment with curative intent.

Survival Outcomes

Survival times varied considerably, with a mean survival duration of 331.76 ± 387.58 days. By the conclusion of the study, 74% of patients had experienced death, underscoring the severity and poor prognosis of advanced lung cancer stages.

Relative Risk (RR)

The analysis of relative risks revealed several important insights regarding factors associated with mortality risk. Male patients exhibit a 14% higher risk of death compared to females . Younger patients, specifically those aged ≤ 70 years, have a 7% lower risk of death compared to their older counterparts. Smoking status also plays a significant role, with smokers having an 18% higher risk of death than non-smokers, while abstinent smokers show a slight reduction in risk by 4% than non-abstinent. Surprisingly, Patients with HBV and HCV have lower mortality risks by 10% and 20%, respectively, compared to those without these conditions. This finding contrasts with the general expectation that chronic viral infections might worsen outcomes.

Screening demonstrates its protective benefits, with screened patients exhibiting a 17% lower risk of death. Although alcohol consumption and diabetes have minimal impact on mortality, thrombosis significantly increases the risk by 28% . Tumor size and cancer stage are critical determinants of mortality. Patients with tumor sizes < 50 mm have a 35% lower risk compared to those with tumor size > 50 mm. Advanced cancer stages elevate mortality risk, with intermediate , advanced , and terminal stages associated with 34%, 32%, and 53% higher risks than patients in early stages.

Diffuse and metastatic cancers are also linked to higher mortality. In contrast, curative treatment significantly reduces the risk of death by 52% . Treatment modalities vary in effectiveness, with treatment1 and treatment6 associated with increased mortality risk, whereas treatment3 , treatment4 , and treatment5 show moderate protective effects.

3.1 Graphical Representation of Tumor Size and Survival

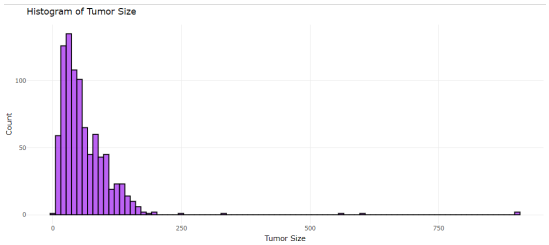

Figure 1: Tumorsize Histogram

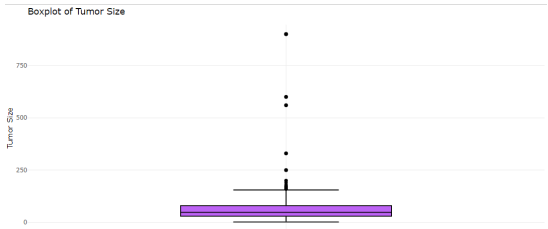

Figure 2: Tumorsize Boxplot

The histogram indicates that the majority of tumors are concentrated in the lower size ranges, with an average tumor size of 31.14 mm. The distribution appears to be right-skewed, as evidenced by the presence of a few larger tumors extending the range up to 900 mm.
The boxplot reveals a median tumor size of 48.00 mm, indicating that half of the tumors are smaller than this value. The interquartile range (IQR), from 30.00 mm (Q1) to 80.00 mm (Q3), highlights the central spread of the data. The upper fence, calculated at 155.00 mm, suggests that tumor sizes beyond this point are considered outliers.

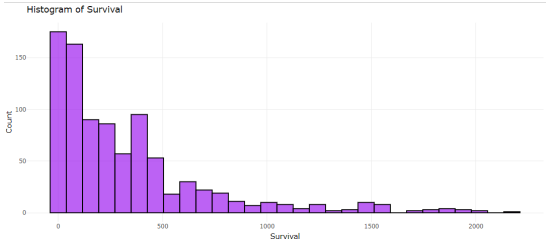

Figure 3: Survival Histogram

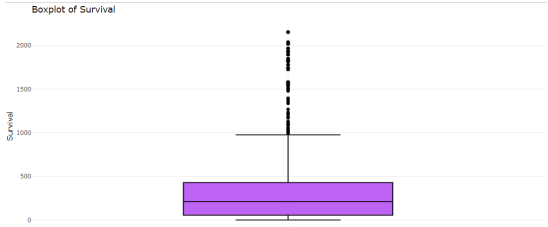

Figure 4: Survival Boxplot

The histogram likely shows a concentration of survival times at the lower end, with fewer observations at higher survival times, indicating a right-skewed distribution.
The boxplot of survival data reveals a median survival time of 211.50 days, indicating that half of the lung cancer patients survived longer than this duration. The Interquartile Range (IQR), from 56.00 days (Q1) to 430.00 days (Q3), shows significant variability: 25% of patients survived less than 56 days, while 75% survived less than 430 days. The upper fence at 97.00 days identifies outliers, suggesting most patients had shorter survival times, with a few experiencing significantly longer survival.

3.2 Mortality And Screening

The mortality rate for patients with lung cancer is estimated at 74.05%. This high mortality rate highlights the severity of lung cancer and emphasizes the urgent need for improvements in early detection and treatment to enhance patient survival.

We also computed a 95% confidence interval for the mortality rate, which is given by [71.18%, 76.92%]. This means that we are 95% confident that the true mortality rate for patients with lung cancer lies within this range, providing a measure of precision for our estimate.

The rate of patients who were screened for cancer is 22.6%. This indicates that only a small portion of the population has undergone screening. Given the importance of early detection in improving cancer outcomes, this relatively low percentage underscores the need for increased efforts to promote and facilitate cancer screening programs.

On the other hand, the rate of unscreened patientss is 77.4%. This high percentage suggests a significant gap in cancer detection, which could lead to late-stage diagnoses. It emphasizes the importance of enhancing access to and awareness of screening services to identify at-risk individuals and improve early detection and treatment opportunities.

3.3 Logistic Regression: Impact of TumorSize on DeathStatus

The logistic regression model investigates the relationship between Tumor Size and the likelihood of death (DeathStatus). The model results show that:

- The coefficient for Tumor Size is -0.0106 (p-value = 5.14e-06), indicating a significant negative relationship between Tumor Size and the likelihood of death. As Tumor Size increases, the likelihood of death decreases slightly.

This result appears to be illogical, as larger tumor sizes are typically expected to correlate with a higher risk of death due to increased disease severity.

3.4 Linear Regression: Impact of TumorSize on SurvivalDays

The linear regression model explores the relationship between Tumor Size and Survival Days. The results show that:

- The coefficient for Tumor Size is -0.9533 (p-value = 6.5e-06), indicating that for every unit (mm) increase in Tumor Size , the number of days survived decreases by approximately 0.9533 days.

The model's R-squared value is 0.02255, suggesting that Tumor Size explains only a small proportion of the variance in survival days. The F-statistic of 20.58 (p-value = 6.499e-06) indicates that the model is statistically significant.

3.5 Estimate Leadtime Bias and Sojourn Time

For doubling time (DT), tumor size of screened patients (DS), and tumor size of non-screened patients (DNS), we fitted a lognormal distribution in each case based on a combination of p-values and AIC values. For DT, all tested distributions had p-values> 0.05, except for the exponential distribution, which had a p-value < 0.05. The lognormal distribution demonstrated the lowest AIC value (1492.557), making it the best fit for DT. Similarly, for DNS, all distributions had p-values > 0.05, and the lognormal distribution showed the lowest AIC value (6944.956), validating it as the best fit. For DS, the lognormal distribution not only had the highest p-value (0.352) but also the lowest AIC value (1835.411), confirming it as the optimal choice. Thus, the lognormal distribution consistently provides the most appropriate fit for these datasets.

The simulation results suggest that the doubling time of tumors varies significantly, with the median doubling time around 96.83 days and a wide range from 11.00 to 438.17 days. This spread indicates that while some tumors double rapidly, others progress much

more slowly. The variability reflects the diversity in tumor behavior among patients, with some tumors showing aggressive growth, while others remain indolent.

The simulated tumor sizes for screened patients have a median value of 37.93 days, with values ranging from 10.21 to 99.46. This distribution reflects the size of tumors at the point of screening detection. Smaller values closer to the minimum suggest earlier-stage tumors, likely identified through screening, while the larger values indicate tumors that may have been detected at later stages but were still within the scope of screening programs.

The simulated tumor sizes for non-screened patients show a median of 50.39, with a range from 10.04 to 196.36. Compared to screened patients, the values for non-screened individuals are generally larger, reflecting the delayed detection of tumors in this group. The broader range suggests that tumors in non-screened patients are diagnosed at more advanced stages, often resulting in larger sizes at the time of diagnosis.

When comparing the lead time bias estimates from Schwartz's method and Duffy's method, we observe distinct differences in the results. According to Schwartz's method for estimating lead time bias, the mean lead time is 304.49 days, with the median at approximately 159 days. This suggests that early detection through screening advances diagnosis by a substantial amount in most cases. However, the wide spread of lead-time values, ranging from 0.76 to 2991.79 days, indicates significant variability in how early detection impacts different patients. In most cases, screening allows for earlier diagnosis, but for some patients, the lead time is minimal or extremely large.

In contrast, Duffy's method estimates a mean lead time of 135.77 days, which is considerably lower than the estimate from Schwartz's method. This means that, according to Duffy's approach, the average overestimation of survival time due to early detection is smaller. After adjusting for lead time bias, the corrected survival time based on Duffy's method is 195.997 days, indicating that the actual survival time, after accounting for the early detection effect, is roughly 196 days.

The contrast between the two methods highlights the significant impact that the choice of method can have on the estimation of lead time bias. Schwartz's method suggests a larger lead time bias, with a higher mean and a wider range of values, while Duffy's method estimates a smaller lead time bias, with a more concentrated range of values. **The distribution that best fits the sojourn time** is the Gamma distribution, which has a p-value of 0.528 and an AIC of 15034.23. Based on these fit statistics, the Gamma distribution is the most appropriate model for the sojourn time, as it minimizes the AIC and provides a reasonably high p-value, indicating a good fit for the data. **The mean sojourn time** (MST) is estimated to be 720.1 days, which represents the average duration from when a disease becomes detectable through screening until it would have naturally progressed to clinical detection without screening. This relatively long MST indicates that, on average, patients remain in a screen-detectable state for a considerable time before clinical symptoms appear. This extended period provides a significant window of opportunity for early detection through screening, making screening programs more effective.

The transition rate, calculated as the reciprocal of the mean sojourn time, is approximately 0.00139 per day. This low rate indicates the slow progression of the disease from its screen-detectable stage to clinical detection. A slower transition rate is beneficial for screening, as it provides more time to diagnose the disease early, before symptoms develop. This further highlights the critical role of screening programs in facilitating early diagnosis and improving outcomes in such cases.

3.6 Kaplan-Meier and Log-Rank Test Results

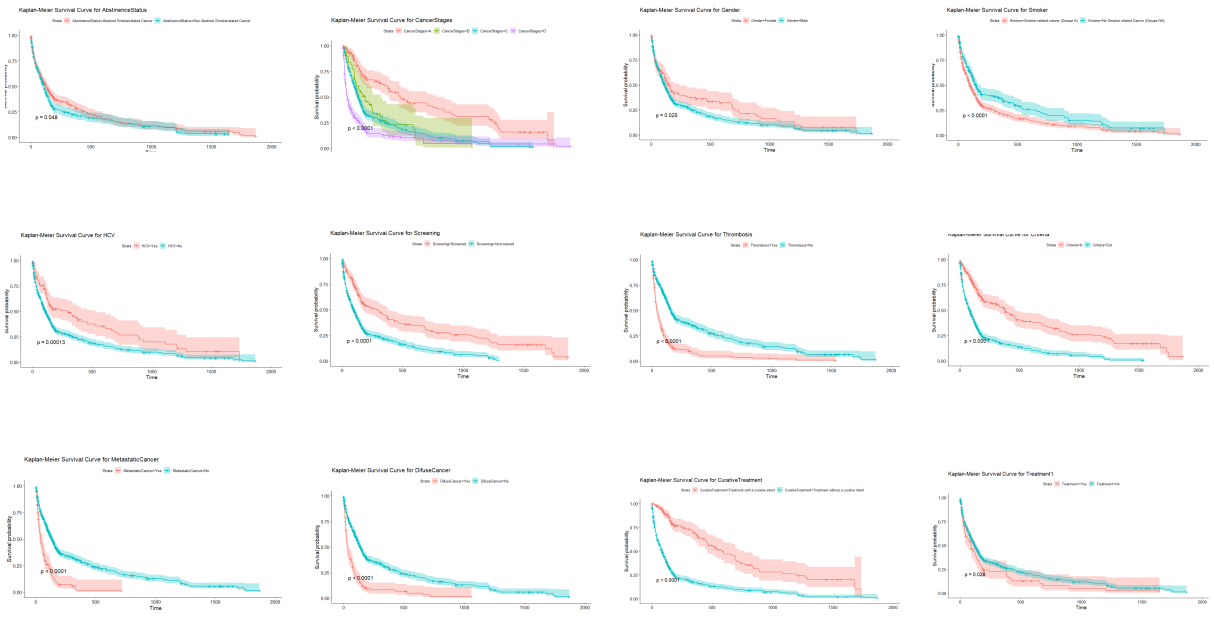

Figure 5: Survival Curves Showing Significant Differences Between Levels of Factors

The log-rank test identified several factors significantly associated with survival in lung cancer patients. Gender showed a significant difference ($p = 0.02$), with males exhibiting worse survival outcomes than females. Smoking status ($p = 6 \times 10^{-5}$) and abstinence status ($p = 0.05$) were also significantly associated with survival, as smokers and non-abstinent patients experienced poorer outcomes.

Hepatitis C (HCV) ($p = 1 \times 10^{-4}$) and screening status ($p = 2 \times 10^{-14}$) were significant predictors, with unscreened patients and those without HCV demonstrating worse survival. Interestingly, patients with HCV had better survival outcomes, a finding that warrants further investigation to understand potential confounding factors or underlying mechanisms.

Thrombosis, cancer staging, diffuse cancer, metastatic cancer, and curative treatment intent were all highly significant ($p < 2 \times 10^{-16}$), with terminal stages and lack of curative treatment strongly associated with poor survival outcomes.

Additionally, Treatment 1 ($p = 0.03$) and criteria ($p < 2 \times 10^{-16}$) were significant. Patients not receiving Treatment 1 had better survival outcomes, possibly reflecting its use in more advanced or severe cases. while those with tumor sizes less than 50 mm exhibited significantly better survival rates.

These findings highlight the importance of early detection, effective screening, and targeted interventions for high-risk groups to improve survival in lung cancer patients.

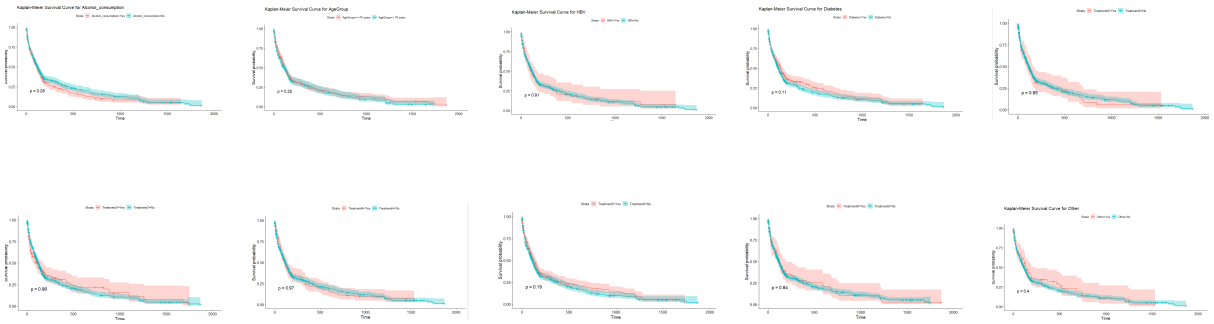

Figure 6: Factors with non-significant differences in survival curve.

Several factors were not significantly associated with survival outcomes in lung cancer patients. Age group ($p = 0.3$), hepatitis B (HBV) ($p = 0.9$), alcohol consumption ($p = 0.3$), and diabetes ($p = 0.1$) did not show a statistically significant impact on survival. Additionally, Treatment 2 ($p = 0.8$), Treatment 3 ($p = 1$), Treatment 4 ($p = 1$), Treatment 5 ($p = 0.2$), and Treatment 6 ($p = 0.8$) were not significantly associated with survival outcomes. These results suggest that these factors may not play a major role in influencing survival in this patient population.

3.7 Cox proportional hazards regression model

Variable	**Hazard Ratio (exp(coef))**	**p-value**
SmokerNo	0.778	0.00564
ScreeningUnscreened	1.414	0.00111
TumorSize	1.001	0.00704
Alcohol$_c$*onsumptionNo*	1.18	0.05750
ThrombosisNo	0.661	5.89×10^{-6}
CriteriaOut	1.749	9.72×10^{-7}
CancerStagesB	0.833	0.39825
CancerStagesC	1.123	0.41094
CancerStagesD	1.910	9.23×10^{-6}
DifuseCancerNo	0.632	2.21×10^{-5}
MetastaticCancerNo	0.705	0.00288
CurativeTreatmentTreatment without a curative intent	2.585	8.04×10^{-13}
Treatment1No	0.737	0.01451

Table 3: Significant Variables in the Cox Proportional Hazards Model

The results of the Cox proportional hazards regression model, obtained through backward selection, provide insights into the factors influencing survival in lung cancer patients. Non-smokers exhibit a reduced risk of death compared to smokers, with a 22% lower risk of death. Unscreened patients show an increased risk of death, being 1.4 times more likely to die than those who were screened. Tumor size plays a critical role, where a slight increase in tumor size corresponds to a marginal increase in the risk of death. The absence of thrombosis reduces the risk of death, with patients without thrombosis having a 34% lower risk of death. Treatment without curative intent significantly increases the risk of death, with those receiving non-curative treatment being 2.58 times more likely to die compared to those receiving curative treatment, highlighting the importance of curative treatment in improving survival outcomes. For alcohol consumption, individuals who do not consume alcohol have a 1.18 times higher risk of death compared to those

who consume alcohol, this finding is unexpected and may reflect underlying confounding factors rather than a direct protective effect of alcohol. Additionally, factors such as cancer stages, diffuse cancer, and metastatic cancer also significantly affect survival. The overall model is highly significant, with a concordance value of 0.753, indicating it effectively ranks survival times for most paired observations. The likelihood ratio test, Wald test, and score (log-rank) test all yield p-values less than 2×10^{-16}, confirming the model's statistical significance.

4 Conclusion

This study highlights the urgent need for early lung cancer detection and intervention. Smoking, advanced cancer stages, and delayed screening significantly increase mortality risks, while early diagnosis and curative treatments offer a lifeline for patients. With a corrected survival time of 195.997 days, the findings underscore the devastating consequences of late detection and inadequate treatment. To reduce mortality and improve survival outcomes, it is imperative to prioritize smoking cessation, expand screening programs, and ensure accessible, effective treatments. While these insights pave the way for life-saving strategies, further research is essential to bridge existing knowledge gaps, refine predictive models, and shape the future of lung cancer care.

References

[1] GeeksforGeeks. How to use the triangular distribution in r, 2023.

[2] Stan Development Team. *RStanArm: Bayesian Applied Regression Modeling via Stan*, 2023.

[3] Stef van Buuren and Karin Groothuis-Oudshoorn. *MICE: Multivariate Imputation by Chained Equations in R*, 2023.

A Appendix

```r
library(readxl)
library(dplyr)
library(ggplot2)
library(survival)
library(survminer)
library(fitdistrplus)
library(triangle)
library(purrr)
library(BaylorEdPsych)
library(mvnmle)
library(mice)
library(Amelia)
library(tidyr)
library(rstanarm)
library(plotly)
library(ggplot2)
library(gridExtra)
library(MASS)
setwd("C:/Users/USER/OneDrive/Desktop/STAT504-AM/PROJECTS")
projdata <-
↪  read_excel("C:\\Users\\USER\\OneDrive\\Desktop\\STAT504-AM\\PROJECTS\\PROJECT
↪  1\\DataProject.xlsx")
to_factor <- c("Gender", "AgeGroup", "Other", "Smoker", "AbstinenceStatus", "HBV",
↪  "HCV",
               "Screening", "Diabetes", "Alcohol_consumption", "Criteria",
               "DifuseCancer", "MetastaticCancer", "DeathStatus", "CancerStages",
               "CurativeTreatment","Thrombosis", "Treatment1", "Treatment2",
               ↪  "Treatment3", "Treatment4", "Treatment5", "Treatment6" )
projdata[to_factor] <- lapply(projdata[to_factor], factor)

projdata$Smoker <- factor(projdata$Smoker, levels = c("Smoker related cancer (Groupe
↪  A)",
"No Smoker related Cancer (Groupe NA)"))
projdata[] <- lapply(projdata, function(x) {
  if (is.factor(x) && all(c("Yes", "No") %in% levels(x))) {
    factor(x, levels = c("Yes", "No"))
  } else {
    x
  }
})

#check the % of missingness in each var
missingness_summary <- sapply(projdata, function(x) sum(is.na(x)) / length(x) * 100)


categorical_vars <- projdata_imputed[, sapply(projdata_imputed, is.factor)]


compute_cramers_v <- function(data) {
  var_names <- colnames(data)
  result <- data.frame(var1 = character(0), var2 = character(0), cramer_v =
  ↪  numeric(0))

  for (i in 1:(ncol(data) - 1)) {
    for (j in (i + 1):ncol(data)) {
```

```
      # Compute Cramér's V for each pair
      cramer_v_val <- assocstats(table(data[, i], data[, j]))$cramer
      result <- rbind(result, data.frame(var1 = var_names[i], var2 = var_names[j],
      ↪  cramer_v = cramer_v_val))
    }
  }

  return(result)
}


cramers_v_matrix <- compute_cramers_v(categorical_vars)


sorted_cramers_v <- cramers_v_matrix[order(-cramers_v_matrix$cramer_v), ]
print(sorted_cramers_v)
#no high correlation between variables

#check the nature of missingness
projdata$missing_SurvivalInDays <- is.na(projdata$SurvivalInDays)
projdata$missing_AgeGroup <- is.na(projdata$AgeGroup)
projdata$missing_AbstinenceStatus <- is.na(projdata$AbstinenceStatus)
projdata$missing_HBV <- is.na(projdata$HBV)
projdata$missing_HCV <- is.na(projdata$HCV)
projdata$missing_Other <- is.na(projdata$Other)
projdata$missing_Screening <- is.na(projdata$Screening)
projdata$missing_TumorSize <- is.na(projdata$TumorSize)
projdata$missing_Diabetes <- is.na(projdata$Diabetes)
projdata$missing_Alcohol_consumption <- is.na(projdata$Alcohol_consumption)
projdata$missing_CancerStages <- is.na(projdata$CancerStages)
projdata$missing_Thrombosis <- is.na(projdata$Thrombosis)
projdata$missing_Criteria <- is.na(projdata$Criteria)
projdata$missing_DifuseCancer <- is.na(projdata$DifuseCancer)
projdata$missing_MetastaticCancer <- is.na(projdata$MetastaticCancer)
projdata$missing_CurativeTreatment <- is.na(projdata$CurativeTreatment)
projdata$missing_DeathStatus <- is.na(projdata$DeathStatus)
projdata$missing_Treatment1 <- is.na(projdata$Treatment1)
projdata$missing_Treatment2 <- is.na(projdata$Treatment2)
projdata$missing_Treatment3 <- is.na(projdata$Treatment3)
projdata$missing_Treatment4 <- is.na(projdata$Treatment4)
projdata$missing_Treatment5 <- is.na(projdata$Treatment5)
projdata$missing_Treatment6 <- is.na(projdata$Treatment6)


imputed_data <- mice(projdata, m = 5, method = 'pmm', maxit = 50)
projdata_imputed <- complete(imputed_data)

model_survival <- glm(missing_SurvivalInDays ~ Gender + Age + AgeGroup + Smoker +
↪  AbstinenceStatus +
                        HBV + HCV + Other + Screening + TumorSize + Diabetes +
                        ↪  Alcohol_consumption +
                        Thrombosis + Criteria + CancerStages + DifuseCancer +
                        ↪  MetastaticCancer +
                        CurativeTreatment + Treatment1 + Treatment2 + Treatment3 +
                        ↪  Treatment4 +
                        Treatment5 + Treatment6 + DeathStatus,
                      data = projdata_imputed, family = binomial)
```

```
model_agegroup <- glm(missing_AgeGroup ~ Gender +
                        SsurvivalInDays + Smoker + AbstinenceStatus +
                         HBV + HCV + Other + Screening + TumorSize + Diabetes +
                         ↪  Alcohol_consumption +
                         Thrombosis + Criteria + CancerStages + DifuseCancer +
                         ↪  MetastaticCancer +
                         CurativeTreatment + Treatment1 + Treatment2 + Treatment3 +
                         ↪  Treatment4 +
                         Treatment5 + Treatment6 + DeathStatus,
                       data = projdata_imputed, family = binomial)
model_abstinence <- glm(missing_AbstinenceStatus ~ Gender +
                          SurvivalInDays + Smoker + AgeGroup +
                          HBV + HCV + Other + Screening + TumorSize + Diabetes +
                          ↪  Alcohol_consumption +
                          Thrombosis + Criteria + CancerStages + DifuseCancer +
                          ↪  MetastaticCancer +
                          CurativeTreatment + Treatment1 + Treatment2 + Treatment3 +
                          ↪  Treatment4 +
                          Treatment5 + Treatment6 + DeathStatus,
                        data = projdata_imputed, family = binomial)
model_hbv <- glm(missing_HBV ~ Gender +
                   SurvivalInDays + Smoker + AgeGroup +
                   AbstinenceStatus + HCV + Other + Screening + TumorSize + Diabetes
                   ↪  + Alcohol_consumption +
                   Thrombosis + Criteria + CancerStages + DifuseCancer +
                   ↪  MetastaticCancer +
                   CurativeTreatment + Treatment1 + Treatment2 + Treatment3 +
                   ↪  Treatment4 +
                   Treatment5 + Treatment6 + DeathStatus,
                 data = projdata_imputed, family = binomial)
model_hcv <- glm(missing_HCV ~ Gender +
                   SurvivalInDays + Smoker + AgeGroup +
                   AbstinenceStatus + HBV + Other + Screening + TumorSize + Diabetes
                   ↪  + Alcohol_consumption +
                   Thrombosis + Criteria + CancerStages + DifuseCancer +
                   ↪  MetastaticCancer +
                   CurativeTreatment + Treatment1 + Treatment2 + Treatment3 +
                   ↪  Treatment4 +
                   Treatment5 + Treatment6 + DeathStatus,
                 data = projdata_imputed, family = binomial)
model_other <- glm(missing_Other ~ Gender +
                     SurvivalInDays + Smoker + AgeGroup +
                     AbstinenceStatus + HBV + HCV + Screening + TumorSize + Diabetes
                     ↪  + Alcohol_consumption +
                     Thrombosis + Criteria + CancerStages + DifuseCancer +
                     ↪  MetastaticCancer +
                     CurativeTreatment + Treatment1 + Treatment2 + Treatment3 +
                     ↪  Treatment4 +
                     Treatment5 + Treatment6 + DeathStatus,
                   data = projdata_imputed, family = binomial)
model_screening <- glm(missing_Screening ~ Gender +
                         SurvivalInDays + Smoker + AgeGroup +
                         AbstinenceStatus + HBV + HCV + Other + TumorSize + Diabetes
                         ↪  + Alcohol_consumption +
                         Thrombosis + Criteria + CancerStages + DifuseCancer +
                         ↪  MetastaticCancer +
                         CurativeTreatment + Treatment1 + Treatment2 + Treatment3 +
                         ↪  Treatment4 +
```

```
                        Treatment5 + Treatment6 + DeathStatus,
                      data = projdata_imputed, family = binomial)
model_tumor <- glm(missing_TumorSize ~ Gender +
                     SurvivalInDays + Smoker + AgeGroup +
                     AbstinenceStatus + HBV + HCV + Other + Screening + Diabetes +
                     ↪  Alcohol_consumption +
                     Thrombosis + Criteria + CancerStages + DifuseCancer +
                     ↪  MetastaticCancer +
                     CurativeTreatment + Treatment1 + Treatment2 + Treatment3 +
                     ↪  Treatment4 +
                     Treatment5 + Treatment6 + DeathStatus,
                   data = projdata_imputed, family = binomial)
model_diabetes <- glm(missing_Diabetes ~ Gender +
                        SurvivalInDays + Smoker + AgeGroup +
                        AbstinenceStatus + HBV + HCV + Other + Screening + TumorSize
                        ↪  + Alcohol_consumption +
                        Thrombosis + Criteria + CancerStages + DifuseCancer +
                        ↪  MetastaticCancer +
                        CurativeTreatment + Treatment1 + Treatment2 + Treatment3 +
                        ↪  Treatment4 +
                        Treatment5 + Treatment6 + DeathStatus,
                      data = projdata_imputed, family = binomial)
model_alcohol <- glm(missing_Alcohol_consumption ~ Gender +
                       SurvivalInDays + Smoker + AgeGroup +
                       AbstinenceStatus + HBV + HCV + Other + Screening + TumorSize +
                       ↪  Diabetes +
                       Thrombosis + Criteria + CancerStages + DifuseCancer +
                       ↪  MetastaticCancer +
                       CurativeTreatment + Treatment1 + Treatment2 + Treatment3 +
                       ↪  Treatment4 +
                       Treatment5 + Treatment6 + DeathStatus,
                     data = projdata_imputed, family = binomial)
model_cancer <- glm(missing_CancerStages ~ Gender +
                      SurvivalInDays + Smoker + AgeGroup +
                      AbstinenceStatus + HBV + HCV + Other + Screening + TumorSize +
                      ↪  Diabetes +
                      Thrombosis + Criteria + Alcohol_consumption + DifuseCancer +
                      ↪  MetastaticCancer +
                      CurativeTreatment + Treatment1 + Treatment2 + Treatment3 +
                      ↪  Treatment4 +
                      Treatment5 + Treatment6 + DeathStatus,
                    data = projdata_imputed, family = binomial)
model_thrombosis <- glm(missing_Thrombosis ~ Gender +
                          SurvivalInDays + Smoker + AgeGroup +
                          AbstinenceStatus + HBV + HCV + Other + Screening +
                          ↪  TumorSize + Diabetes +
                          CancerStages + Criteria + Alcohol_consumption +
                          ↪  DifuseCancer + MetastaticCancer +
                          CurativeTreatment + Treatment1 + Treatment2 + Treatment3 +
                          ↪  Treatment4 +
                          Treatment5 + Treatment6 + DeathStatus,
                        data = projdata_imputed, family = binomial)
model_criteria <- glm(missing_Criteria ~ Gender +
                        SurvivalInDays + Smoker + AgeGroup +
                        AbstinenceStatus + HBV + HCV + Other + Screening + TumorSize
                        ↪  + Diabetes +
                        CancerStages + Thrombosis + Alcohol_consumption +
                        ↪  DifuseCancer + MetastaticCancer +
```

```
                        CurativeTreatment + Treatment1 + Treatment2 + Treatment3 +
                        ↪  Treatment4 +
                        Treatment5 + Treatment6 + DeathStatus,
                      data = projdata_imputed, family = binomial)
model_diffuse <- glm(missing_DifuseCancer ~ Gender +
                       SurvivalInDays + Smoker + AgeGroup +
                       AbstinenceStatus + HBV + HCV + Other + Screening + TumorSize +
                       ↪  Diabetes +
                       CancerStages + Thrombosis + Alcohol_consumption + Criteria +
                       ↪  MetastaticCancer +
                       CurativeTreatment + Treatment1 + Treatment2 + Treatment3 +
                       ↪  Treatment4 +
                       Treatment5 + Treatment6 + DeathStatus,
                     data = projdata_imputed, family = binomial)
model_metastatic <- glm(missing_MetastaticCancer ~ Gender +
                          SurvivalInDays + Smoker + AgeGroup +
                          AbstinenceStatus + HBV + HCV + Other + Screening +
                          ↪  TumorSize + Diabetes +
                          CancerStages + Thrombosis + Alcohol_consumption + Criteria
                          ↪  + DifuseCancer +
                          CurativeTreatment + Treatment1 + Treatment2 + Treatment3 +
                          ↪  Treatment4 +
                          Treatment5 + Treatment6 + DeathStatus,
                        data = projdata_imputed, family = binomial)
model_curative <- glm(missing_CurativeTreatment ~ Gender +
                        SurvivalInDays + Smoker + AgeGroup +
                        AbstinenceStatus + HBV + HCV + Other + Screening + TumorSize
                        ↪  + Diabetes +
                        CancerStages + Thrombosis + Alcohol_consumption + Criteria +
                        ↪  DifuseCancer +
                        MetastaticCancer + Treatment1 + Treatment2 + Treatment3 +
                        ↪  Treatment4 +
                        Treatment5 + Treatment6 + DeathStatus,
                      data = projdata_imputed, family = binomial)
model_death <- glm(missing_DeathStatus ~ Gender +
                     SurvivalInDays + Smoker + AgeGroup +
                     AbstinenceStatus + HBV + HCV + Other + Screening + TumorSize +
                     ↪  Diabetes +
                     CancerStages + Thrombosis + Alcohol_consumption + Criteria +
                     ↪  DifuseCancer +
                     MetastaticCancer + Treatment1 + Treatment2 + Treatment3 +
                     ↪  Treatment4 +
                     Treatment5 + Treatment6 + CurativeTreatment,
                   data = projdata_imputed, family = binomial)
model_treatment1 <- glm(missing_Treatment1 ~ Gender +
                          SurvivalInDays + Smoker + AgeGroup +
                          AbstinenceStatus + HBV + HCV + Other + Screening +
                          ↪  TumorSize + Diabetes +
                          CancerStages + Thrombosis + Alcohol_consumption + Criteria
                          ↪  + DifuseCancer +
                          MetastaticCancer + DeathStatus + Treatment2 + Treatment3 +
                          ↪  Treatment4 +
                          Treatment5 + Treatment6 + CurativeTreatment,
                        data = projdata_imputed, family = binomial)
model_treatment2 <- glm(missing_Treatment2 ~ Gender +
                          SurvivalInDays + Smoker + AgeGroup +
                          AbstinenceStatus + HBV + HCV + Other + Screening +
                          ↪  TumorSize + Diabetes +
```

```
                          CancerStages + Thrombosis + Alcohol_consumption + Criteria
                          ↪  + DifuseCancer +
                          MetastaticCancer + DeathStatus + Treatment1 + Treatment3 +
                          ↪  Treatment4 +
                          Treatment5 + Treatment6 + CurativeTreatment,
                        data = projdata_imputed, family = binomial)
model_treatment3 <- glm(missing_Treatment3 ~ Gender +
                          SurvivalInDays + Smoker + AgeGroup +
                          AbstinenceStatus + HBV + HCV + Other + Screening +
                          ↪  TumorSize + Diabetes +
                          CancerStages + Thrombosis + Alcohol_consumption + Criteria
                          ↪  + DifuseCancer +
                          MetastaticCancer + DeathStatus + Treatment1 + Treatment2 +
                          ↪  Treatment4 +
                          Treatment5 + Treatment6 + CurativeTreatment,
                        data = projdata_imputed, family = binomial)
model_treatment4 <- glm(missing_Treatment4 ~ Gender +
                          SurvivalInDays + Smoker + AgeGroup +
                          AbstinenceStatus + HBV + HCV + Other + Screening +
                          ↪  TumorSize + Diabetes +
                          CancerStages + Thrombosis + Alcohol_consumption + Criteria
                          ↪  + DifuseCancer +
                          MetastaticCancer + DeathStatus + Treatment1 + Treatment2 +
                          ↪  Treatment3 +
                          Treatment5 + Treatment6 + CurativeTreatment,
                        data = projdata_imputed, family = binomial)
model_treatment5 <- glm(missing_Treatment5 ~ Gender +
                          SurvivalInDays + Smoker + AgeGroup +
                          AbstinenceStatus + HBV + HCV + Other + Screening +
                          ↪  TumorSize + Diabetes +
                          CancerStages + Thrombosis + Alcohol_consumption + Criteria
                          ↪  + DifuseCancer +
                          MetastaticCancer + DeathStatus + Treatment1 + Treatment2 +
                          ↪  Treatment3 +
                          Treatment4 + Treatment6 + CurativeTreatment,
                        data = projdata_imputed, family = binomial)
model_treatment6 <- glm(missing_Treatment6 ~  Gender +
                          SurvivalInDays + Smoker + AgeGroup +
                          AbstinenceStatus + HBV + HCV + Other + Screening +
                          ↪  TumorSize + Diabetes +
                          CancerStages + Thrombosis + Alcohol_consumption + Criteria
                          ↪  + DifuseCancer +
                          MetastaticCancer + DeathStatus + Treatment1 + Treatment2 +
                          ↪  Treatment3 +
                          Treatment4 + Treatment5 + CurativeTreatment,
                        data = projdata_imputed, family = binomial)
#all factors are MAR but(agegroup, abstinencestatus, criteria, difusecancer)



mar_vars <- c("SurvivalInDays", "HBV", "HCV", "Other", "Screening", "TumorSize",
              "Diabetes", "Alcohol_consumption", "CancerStages", "Thrombosis",
              "MetastaticCancer", "CurativeTreatment", "DeathStatus",
              "Treatment1", "Treatment2", "Treatment3", "Treatment4", "Treatment5",
              ↪  "Treatment6")

# Subset the dataset to only include the MAR variables
```

```r
mar_data <- projdata[, mar_vars]


quant_vars <- c("SurvivalInDays", "TumorSize")
quant_data <- mar_data[, quant_vars]
cat_vars <- setdiff(mar_vars, quant_vars)


quant_imputed <- mice(quant_data, m = 5, method = "pmm", seed = 500)


completed_quant_data <- complete(quant_imputed, 1)


mar_data[, quant_vars] <- completed_quant_data
projdata[, quant_vars] <- mar_data[, quant_vars]

cancer_stage_data <- mar_data[, "CancerStages", drop = FALSE]


cancer_stage_data$Dummy <- 1

cancer_stage_imputed <- mice(cancer_stage_data, m = 5, method = "polyreg", seed =
↪  500)


completed_cancer_stage_data <- complete(cancer_stage_imputed, 1)


completed_cancer_stage_data$Dummy <- NULL


mar_data[, "CancerStages"] <- completed_cancer_stage_data[, "CancerStages"]
projdata[, "CancerStages"] <- mar_data[, "CancerStages"]


binary_cat_vars <- setdiff(cat_vars, "CancerStages")
binary_cat_data <- mar_data[, binary_cat_vars]


binary_cat_imputed <- mice(binary_cat_data, m = 5, method = "logreg", seed = 500)


completed_binary_cat_data <- complete(binary_cat_imputed, 1)


mar_data[, binary_cat_vars] <- completed_binary_cat_data
projdata[, binary_cat_vars] <- mar_data[, binary_cat_vars]


mcar_data_full <- projdata[, c("AbstinenceStatus", "Criteria", "AgeGroup", "Age",
↪  "DifuseCancer"), drop = FALSE]


mcar_result_full <- LittleMCAR(mcar_data_full)

mcar_result_full
```

```r
#missingness is NOT MCAR
#IMPUTATION FOR MNAR MISSING

formula <- AbstinenceStatus ~ SurvivalInDays + HBV + HCV + Other + Screening +
↪  TumorSize +
  Diabetes + Alcohol_consumption + CancerStages + Thrombosis + MetastaticCancer +
  CurativeTreatment + DeathStatus + Treatment1 + Treatment2 + Treatment3 +
  Treatment4 + Treatment5 + Treatment6

model <- stan_glm(formula,
                  data = projdata,
                  family = binomial,
                  prior = normal(0, 1),  # Prior for coefficients
                  prior_intercept = normal(0, 5),  # Prior for intercept
                  seed = 500)


predicted_missing_probabilities <- predict(model, type = "response")
predicted_missing_probabilities[is.na(predicted_missing_probabilities)] <- 0.5


missing_indices <- is.na(projdata$AbstinenceStatus)
imputed_values <- rbinom(sum(missing_indices), 1,
↪  predicted_missing_probabilities[missing_indices])
projdata$AbstinenceStatus[missing_indices] <- factor(imputed_values, levels = c(0,
↪  1), labels = levels(projdata$AbstinenceStatus))


formula_age_group <- AgeGroup ~ SurvivalInDays + HBV + HCV + Other + Screening +
↪  TumorSize +
  Diabetes + Alcohol_consumption + CancerStages + Thrombosis + MetastaticCancer +
  CurativeTreatment + DeathStatus + Treatment1 + Treatment2 + Treatment3 +
  Treatment4 + Treatment5 + Treatment6
model_age_group <- stan_glm(formula_age_group,
                            data = projdata,
                            family = binomial,  # If AgeGroup has multiple
                            ↪  categories
                            prior = normal(0, 1),
                            prior_intercept = normal(0, 5),
                            seed = 500)


predicted_missing_probabilities_age_group <- predict(model_age_group, type =
↪  "response")
predicted_missing_probabilities_age_group[is.na(predicted_missing_probabilities_age_group)]
↪  <- 0.5


missing_indices_age_group <- is.na(projdata$AgeGroup)
imputed_values_age_group <- rbinom(sum(missing_indices_age_group), 1,
predicted_missing_probabilities_age_group[missing_indices_age_group])
projdata$AgeGroup[missing_indices_age_group] <- factor(imputed_values_age_group,
↪  levels = c(0, 1),  labels = levels(projdata$AgeGroup))

formula_difuse_cancer <- DifuseCancer ~ SurvivalInDays + HBV + HCV + Other +
↪  Screening + TumorSize +
  Diabetes + Alcohol_consumption + CancerStages + Thrombosis + MetastaticCancer +
```

```
  CurativeTreatment + DeathStatus + Treatment1 + Treatment2 + Treatment3 +
  Treatment4 + Treatment5 + Treatment6
model_difuse_cancer <- stan_glm(formula_difuse_cancer,
                                data = projdata,
                                family = binomial,
                                prior = normal(0, 1),
                                prior_intercept = normal(0, 5),
                                seed = 500)


predicted_missing_probabilities_difuse_cancer <- predict(model_difuse_cancer, type =
↪  "response")
predicted_missing_probabilities_difuse_cancer[is.na(predicted_missing_probabilities_difuse_cancer)]
↪  <- 0.5


missing_indices_difuse_cancer <- is.na(projdata$DifuseCancer)
imputed_values_difuse_cancer <-rbinom(sum(missing_indices_difuse_cancer), 1,
predicted_missing_probabilities_difuse_cancer[missing_indices_difuse_cancer])
projdata$DifuseCancer[missing_indices_difuse_cancer] <-
↪  factor(imputed_values_difuse_cancer,  levels = c(0, 1),
labels = levels(projdata$DifuseCancer))


formula_criteria <- Criteria ~ AbstinenceStatus + AgeGroup + Age + DifuseCancer +
  HBV + HCV + Other + Screening + TumorSize + Diabetes + Alcohol_consumption +
  CancerStages + Thrombosis + MetastaticCancer + CurativeTreatment + DeathStatus +
  Treatment1 + Treatment2 + Treatment3 + Treatment4 + Treatment5 + Treatment6


model_criteria <- stan_glm(formula_criteria,
                           data = projdata,
                           family = "binomial",  # or "multinomial" if Criteria has
                           ↪  multiple levels
                           prior = normal(0, 1),
                           prior_intercept = normal(0, 5),
                           seed = 500)


predicted_missing_probabilities_criteria <- predict(model_criteria, type =
↪  "response")


predicted_missing_probabilities_criteria[is.na(predicted_missing_probabilities_criteria)]
↪  <- 0.5


missing_indices_criteria <- is.na(projdata$Criteria)
imputed_values_criteria <- rbinom(sum(missing_indices_criteria), 1,
↪  predicted_missing_probabilities_criteria[missing_indices_criteria])


projdata$Criteria[missing_indices_criteria] <- factor(imputed_values_criteria,
↪  levels = levels(projdata$Criteria))
```

```r
imputed <- mice(projdata[, c("AgeGroup", "Criteria", "AbstinenceStatus")], method =
↪  "logreg", m = 5, seed = 500)


completed <- complete(imputed, 1)  # Get the first imputed dataset


projdata$AgeGroup <- completed$AgeGroup
projdata$Criteria <- completed$Criteria
projdata$AbstinenceStatus <- completed$AbstinenceStatus

summary(projdata$AgeGroup)
summary(projdata$Criteria)
summary(projdata$AbstinenceStatus)

library(randomForest)

missing_indices_age <- is.na(projdata$Age)

rf_model <- randomForest(Age ~ AgeGroup, data = projdata, na.action = na.exclude)


predicted_age_rf <- predict(rf_model, newdata = projdata)


projdata$Age[missing_indices_age] <- predicted_age_rf[missing_indices_age]


summary(projdata$Age)


summary_qual <- projdata %>%
  select_if(is.factor) %>%
  gather(key = "Variable", value = "Level") %>%
  count(Variable, Level, name = "Count") %>%
  group_by(Variable) %>%
  mutate(Percentage = Count / sum(Count) * 100) %>%
  ungroup()

summary_qual

summary_quant <- projdata %>%
  summarise(across(
    where(is.numeric) & !starts_with("ID"),
    list(
      mean = ~mean(.),
      sd = ~sd(.),
      ci_lower = ~mean(.) - 1.96 * sd(.) / sqrt(n()),
      ci_upper = ~mean(.) + 1.96 * sd(.) / sqrt(n())
    ),
    .names = "{.col}_{.fn}"
  ))

print(summary_quant)

calculate_binwidth <- function(data) {
  IQR_value <- IQR(data)
```

```
  n <- length(data)
  bin_width <- 2 * IQR_value * (n ^ (-1/3))
  return(bin_width)
}

bin_width <- calculate_binwidth(projdata$TumorSize)
print(paste("Recommended binwidth:", bin_width))

# Graphs
plot_tumor_size_interactive1 <- ggplot(projdata, aes(x = TumorSize)) +
  geom_histogram(binwidth = 10.38, fill = "purple", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Tumor Size", x = "Tumor Size", y = "Count") +
  theme_minimal()

plot_tumor_size_interactive2 <- ggplot(projdata, aes(x = "", y = TumorSize)) +
  geom_boxplot(fill = "purple", color = "black", alpha = 0.7) +
  labs(title = "Boxplot of Tumor Size", x = "", y = "Tumor Size") +
  theme_minimal()

interactive_plot1 <- ggplotly(plot_tumor_size_interactive1)
interactive_plot2 <- ggplotly(plot_tumor_size_interactive2)

plot_SurvivalInDays_interactive1 <- ggplot(projdata, aes(x = SurvivalInDays)) +
  geom_histogram(binwidth = 77.59, fill = "purple", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Survival", x = "Survival", y = "Count") +
  theme_minimal()

plot1 <- ggplotly(plot_SurvivalInDays_interactive1)

plot_SurvivalInDays_interactive2 <- ggplot(projdata, aes(x = "", y = SurvivalInDays))
↪  +
  geom_boxplot(fill = "purple", color = "black", alpha = 0.7) +
  labs(title = "Boxplot of Survival", x = "", y = "Survival") +
  theme_minimal()

plot2 <- ggplotly(plot_SurvivalInDays_interactive2)


deaths_nb <- sum(projdata$DeathStatus == "Death")

population_size <- length(unique(projdata$ID))


mortality_rate <- deaths_nb / population_size
mortality_rate_percent <- mortality_rate * 100

z_score <- 1.96  # for 95% confidence level

ci_lower <- mortality_rate - z_score * sqrt((mortality_rate * (1 - mortality_rate)) /
↪  population_size)
ci_upper <- mortality_rate + z_score * sqrt((mortality_rate * (1 - mortality_rate)) /
↪  population_size)

ci_lower_percent <- ci_lower * 100
ci_upper_percent <- ci_upper * 100

cat("Mortality Rate:", round(mortality_rate_percent, 2), "%\n")
```

```r
cat("95% Confidence Interval: [", round(ci_lower_percent, 2), "%, ",
↪  round(ci_upper_percent, 2), "%]\n")


screened_patient <- sum(projdata$Screening == "Screened")

unscreened_patient <- sum(projdata$Screening == "Unscreened")


screened_rate <- screened_patient / population_size
screened_rate_percent <- screened_rate * 100

unscreened_rate <- unscreened_patient / population_size
unscreened_rate_percent <- unscreened_rate * 100

cat("Rate of Screened Cancer:", round(screened_rate_percent, 2), "%\n")
cat("Rate of Unscreened Cancer:", round(unscreened_rate_percent, 2), "%\n")

calculate_rr_table <- function(variable_name, dataset) {

  levels <- unique(dataset[[variable_name]])


  rr_results <- data.frame(Variable = character(),
                           Level = character(),
                           Risk = numeric(),
                           RR = numeric(),
                           CI_Lower = numeric(),
                           CI_Upper = numeric(),
                           stringsAsFactors = FALSE)


  if (length(levels) > 2) {
    # Set "A" as the reference level (if present)
    reference_level <- "A"
    if (!reference_level %in% levels) {
      stop(paste("Reference level", reference_level, "is not in the levels of",
      ↪  variable_name))
    }


    group_ref <- dataset[dataset[[variable_name]] == reference_level, ]
    risk_ref <- mean(group_ref$DeathStatus == "Death")


    for (level in levels) {

      group <- dataset[dataset[[variable_name]] == level, ]


      risk_group <- mean(group$DeathStatus == "Death")


      if (level == reference_level) {
        rr <- 1
        ci_lower <- 1
        ci_upper <- 1
```

```
      } else {
        rr <- risk_group / risk_ref


        n_ref <- nrow(group_ref)
        n_group <- nrow(group)
        se <- sqrt((1 / (n_ref * risk_ref * (1 - risk_ref))) + (1 / (n_group *
        ↪  risk_group * (1 - risk_group))))


        ci_lower <- exp(log(rr) - 1.96 * se)
        ci_upper <- exp(log(rr) + 1.96 * se)
      }


      rr_results <- rbind(rr_results, data.frame(Variable = variable_name, Level =
      ↪  level, Risk = risk_group, RR = rr, CI_Lower = ci_lower, CI_Upper =
      ↪  ci_upper))
    }
  } else {

    for (level in levels) {

      group <- dataset[dataset[[variable_name]] == level, ]


      risk_group <- mean(group$DeathStatus == "Death")


      other_group <- dataset[dataset[[variable_name]] != level, ]
      risk_other <- mean(other_group$DeathStatus == "Death")


      rr <- risk_group / risk_other


      n_group <- nrow(group)
      n_other_group <- nrow(other_group)
      se <- sqrt((1 / (n_group * risk_group * (1 - risk_group))) + (1 /
      ↪  (n_other_group * risk_other * (1 - risk_other))))


      ci_lower <- exp(log(rr) - 1.96 * se)
      ci_upper <- exp(log(rr) + 1.96 * se)


      rr_results <- rbind(rr_results, data.frame(Variable = variable_name, Level =
      ↪  level, Risk = risk_group, RR = rr, CI_Lower = ci_lower, CI_Upper =
      ↪  ci_upper))
    }
  }

  return(rr_results)
}


categorical_vars <- names(projdata)[sapply(projdata, is.factor)]  # Get column names
↪  where data is of type factor
```

```r
categorical_vars <- categorical_vars[!categorical_vars %in% c("DeathStatus", "ID")]


all_rr_results <- data.frame(Variable = character(),
                             Level = character(),
                             Risk = numeric(),
                             RR = numeric(),
                             CI_Lower = numeric(),
                             CI_Upper = numeric(),
                             stringsAsFactors = FALSE)


for (var in categorical_vars) {
  rr_result <- calculate_rr_table(var, projdata)
  all_rr_results <- rbind(all_rr_results, rr_result)
}


print(all_rr_results)


logistic_model <- glm(DeathStatus ~ TumorSize, data = projdata, family = binomial)

summary(logistic_model)


linear_model <- lm(SurvivalInDays ~ TumorSize, data = projdata)

summary(linear_model)

dbltm <- read.csv("C:\\Users\\USER\\OneDrive\\Desktop\\STAT504-AM\\PROJECTS\\PROJECT
↪  1\\DoublingTime.csv")

dbltm$DT <- dbltm$DT*365.25

hist(dbltm$DT, breaks=30, col="purple", main="Tumor Doubling Time", xlab="Days",
↪  freq=FALSE)

gamma_fit <- fitdistr(dbltm$DT, "gamma")
ks_gamma <- ks.test(dbltm$DT, "pgamma", 1.962174675, 0.015487722)


weibull_fit <- fitdistr(dbltm$DT, "weibull")
ks_weibull <- ks.test(dbltm$DT, "pweibull", 1.43849737, 140.35103769)


exp_fit <- fitdistr(dbltm$DT, "exponential")
ks_exp <- ks.test(dbltm$DT, "pexp", 0.0078931390 )


lognorm_fit <- fitdistr(dbltm$DT, "lognormal")
ks_lognorm <- ks.test(dbltm$DT, "plnorm", 4.56580355, 0.77142052)


ks_results <- data.frame(
  Distribution = c("Gamma", "Weibull", "Exponential", "Lognormal"),
  D_statistic = c(ks_gamma$statistic, ks_weibull$statistic, ks_exp$statistic,
  ↪  ks_lognorm$statistic),
```

```
  P_value = c(ks_gamma$p.value, ks_weibull$p.value, ks_exp$p.value,
  ↪  ks_lognorm$p.value)
)


print(ks_results)


log_likelihoods <- c(logLik(gamma_fit), logLik(weibull_fit), logLik(exp_fit),
↪  logLik(lognorm_fit))

aic_values <- c(AIC(gamma_fit), AIC(weibull_fit), AIC(exp_fit), AIC(lognorm_fit))

logLik_AIC_results <- data.frame(
  Distribution = c("Gamma", "Weibull", "Exponential", "Lognormal"),
  Log_Likelihood = log_likelihoods,
  AIC = aic_values
)

print(logLik_AIC_results)

data1 <- read_excel("C:/Users/USER/OneDrive/Desktop/STAT504-AM/PROJECTS/PROJECT
↪  1/DS.xlsx")
data2 <- read_excel("C:/Users/USER/OneDrive/Desktop/STAT504-AM/PROJECTS/PROJECT
↪  1/DNS.xlsx")


gamma_fit_DS <- fitdistr(data1$DS, "gamma")
gamma_sim_DS <- rgamma(length(data1$DS), 1.970091011, 0.043185846 )
ks_gamma_DS <- ks.test(data1$DS, "pgamma", 1.970091011, 0.043185846 )

weibull_fit_DS <- fitdistr(data1$DS, "weibull")
weibull_sim_DS <- rweibull(length(data1$DS), 1.15700373, 48.73659254)
ks_weibull_DS <- ks.test(data1$DS, "pweibull", 1.15700373, 48.73659254)

exp_fit_DS <- fitdistr(data1$DS, "exponential")
exp_sim_DS <- rexp(length(data1$DS), 0.021920781 )
ks_exp_DS <- ks.test(data1$DS, "pexp", 0.021920781 )

lognorm_fit_DS <- fitdistr(data1$DS, "lognormal")
lognorm_sim_DS <- rlnorm(length(data1$DS), 3.54555220, 0.64966652)
ks_lognorm_DS <- ks.test(data1$DS, "plnorm", 3.54555220, 0.64966652)

ks_results_DS <- data.frame(
  Distribution = c("Gamma", "Weibull", "Exponential", "Lognormal"),
  D_statistic = c(ks_gamma_DS$statistic, ks_weibull_DS$statistic,
  ↪  ks_exp_DS$statistic, ks_lognorm_DS$statistic),
  P_value = c(ks_gamma_DS$p.value, ks_weibull_DS$p.value, ks_exp_DS$p.value,
  ↪  ks_lognorm_DS$p.value)
)

print(ks_results_DS)

log_likelihoods_DS <- c(logLik(gamma_fit_DS), logLik(weibull_fit_DS),
↪  logLik(exp_fit_DS), logLik(lognorm_fit_DS))
```

```r
aic_values_DS <- c(AIC(gamma_fit_DS), AIC(weibull_fit_DS), AIC(exp_fit_DS),
↪  AIC(lognorm_fit_DS))


logLik_AIC_results_DS <- data.frame(
  Distribution = c("Gamma", "Weibull", "Exponential", "Lognormal"),
  Log_Likelihood = log_likelihoods_DS,
  AIC = aic_values_DS
)

print(logLik_AIC_results_DS)


gamma_fit_DNS <- fitdistr(data2$DNS, "gamma")
gamma_sim_DNS <- rgamma(length(data2$DNS), 2.200249921, 0.033368808 )
ks_gamma_DNS <- ks.test(data2$DNS, "pgamma", 2.200249921, 0.033368808 )


weibull_fit_DNS <- fitdistr(data2$DNS, "weibull")
weibull_sim_DNS <- rweibull(length(data2$DNS), 1.38867014, 72.88813196)
ks_weibull_DNS <- ks.test(data2$DNS, "pweibull", 1.38867014, 72.88813196)


exp_fit_DNS <- fitdistr(data2$DNS, "exponential")
exp_sim_DNS <- rexp(length(data2$DNS), 0.0151651290 )
ks_exp_DNS <- ks.test(data2$DNS, "pexp", 0.0151651290 )


lognorm_fit_DNS <- fitdistr(data2$DNS, "lognormal")
lognorm_sim_DNS <- rlnorm(length(data2$DNS), 3.94476185, 0.70570463)
ks_lognorm_DNS <- ks.test(data2$DNS, "plnorm", 3.94476185, 0.70570463)


ks_results_DNS <- data.frame(
  Distribution = c("Gamma", "Weibull", "Exponential", "Lognormal"),
  D_statistic = c(ks_gamma_DNS$statistic, ks_weibull_DNS$statistic,
  ↪  ks_exp_DNS$statistic, ks_lognorm_DNS$statistic),
  P_value = c(ks_gamma_DNS$p.value, ks_weibull_DNS$p.value, ks_exp_DNS$p.value,
  ↪  ks_lognorm_DNS$p.value)
)


print(ks_results_DNS)


log_likelihoods_DNS <- c(logLik(gamma_fit_DNS), logLik(weibull_fit_DNS),
↪  logLik(exp_fit_DNS), logLik(lognorm_fit_DNS))


aic_values_DNS <- c(AIC(gamma_fit_DNS), AIC(weibull_fit_DNS), AIC(exp_fit_DNS),
↪  AIC(lognorm_fit_DNS))


logLik_AIC_results_DNS <- data.frame(
  Distribution = c("Gamma", "Weibull", "Exponential", "Lognormal"),
  Log_Likelihood = log_likelihoods_DNS,
  AIC = aic_values_DNS
```

```
)

print(logLik_AIC_results_DNS)

set.seed(123)
M <- 1000
lognorm_fit_DT <- fitdistr(dbltm$DT, "lognormal")
DT_sim <- rlnorm(M, 4.56580355, 0.77142052)
for (i in 1:M) {
  if (DT_sim [i] > 440) {DT_sim[i]=mean(DT_sim)}
  else {DT_sim[i]=DT_sim[i]}}
summary(DT_sim)

lognorm_fit_DS <- fitdistr(data1$DS, "lognormal")
DS_sim <- rlnorm(M, 3.54555220, 0.64966652)
for (i in 1:M) {
  if (DS_sim[i]> 100||DS_sim[i]< 10) {DS_sim[i]=mean(DS_sim)}
  else {DS_sim[i]=DS_sim[i]}}
summary(DS_sim)

lognorm_fit_DNS <- fitdistr(data2$DNS, "lognormal")
DNS_sim <- rlnorm(M, 3.94476185, 0.70570463 )
for (i in 1:M) {
  if (DNS_sim[i]> 200||DNS_sim[i]< 10) {DNS_sim[i]=mean(DNS_sim)}
  else {DNS_sim[i]=DNS_sim[i]}}
summary(DNS_sim)

LT_sim<- 3*DT_sim*((log(DNS_sim/DS_sim))/log(2))
LT_sim<- replace (LT_sim, LT_sim<0,mean(LT_sim))

summary(LT_sim)

M <- 1000
min_value <- 10
mode_value <- 12
max_value <- 14
z <- rtriangle(n = M, a = min_value, b = max_value,
               c = mode_value)

SojournTime <- 3 * DT_sim * ((log(DNS_sim / z)) / log(2))
SojournTime <- replace(SojournTime, SojournTime < 0, mean(SojournTime))
SojournTime <- replace(SojournTime, SojournTime > 3650, mean(SojournTime))
mean(SojournTime)

t <- 365
n <- 894
A2 <- NULL
B2 <- NULL
LeadTimeM22 <- NULL
SurvivalInDays <- projdata$SurvivalInDays
for (i in 1:n) {
  if (SurvivalInDays[i]<=t)
  {
    A2 [i] <-1-exp(-(1/mean(SojournTime))*SurvivalInDays[i])

 ↪  -(1/mean(SojournTime))*SurvivalInDays[i]*exp(-(1/mean(SojournTime))*SurvivalInDays[i])
    B2 [i] <-(1/mean(SojournTime))*(1-exp(-(1/mean(SojournTime))*SurvivalInDays[i]))
```

```
    LeadTimeM22 [i] <- A2[i]/B2[i]
  }
  else { LeadTimeM22 [i]<- (1-exp(-(1/mean(SojournTime))*t))/(1/mean(SojournTime))}
}

summary(LeadTimeM22)

CorrectedSurvival12<- SurvivalInDays-LeadTimeM22
mean(CorrectedSurvival12)



gamma_fit_ST <- fitdistr(SojournTime, "gamma")
ks_gamma_ST <- ks.test(SojournTime, "pgamma",  gamma_fit_ST$estimate[1],
↪  gamma_fit_ST$estimate[2])


weibull_fit_ST <- fitdistr(SojournTime, "weibull")
ks_weibull_ST <- ks.test(SojournTime, "pweibull", 1.29294923, 784.86070315)


exp_fit_ST <- fitdistr(SojournTime, "exponential")
ks_exp_ST <- ks.test(SojournTime, "pexp", exp_fit_ST$estimate[1] )



ks_results_ST <- data.frame(
  Distribution = c("Gamma", "Weibull", "Exponential"),
  D_statistic = c(ks_gamma_ST$statistic, ks_weibull_ST$statistic,
  ↪  ks_exp_ST$statistic),
  P_value = c(ks_gamma_ST$p.value, ks_weibull_ST$p.value, ks_exp_ST$p.value)
)


print(ks_results_ST)


log_likelihoods_ST <- c(logLik(gamma_fit_ST), logLik(weibull_fit_ST),
↪  logLik(exp_fit_ST))


aic_values_ST <- c(AIC(gamma_fit_ST), AIC(weibull_fit_ST), AIC(exp_fit_ST))


logLik_AIC_results_ST <- data.frame(
  Distribution = c("Gamma", "Weibull", "Exponential"),
  Log_Likelihood = log_likelihoods_ST,
  AIC = aic_values_ST
)

print(logLik_AIC_results_ST)

transition_rate <- 1 /mean(SojournTime)



projdata$CorrectedSurvival12 <- SurvivalInDays - LeadTimeM22
```

```r
projdata$DeathStatus <- ifelse(projdata$DeathStatus == "Death", 1, 0)



factors <- setdiff(names(projdata)[sapply(projdata, is.factor)], "DeathStatus")


for (factor_name in factors) {


  km_fit <- survfit(Surv(CorrectedSurvival12, DeathStatus) ~
  ↪  as.factor(projdata[[factor_name]]), data = projdata)

  # Plot Kaplan-Meier survival curves
  plot_title <- paste("Kaplan-Meier Survival Curves for", factor_name)
  ggsurvplot(km_fit, data = projdata, pval = TRUE, conf.int = TRUE,
             title = plot_title)

  #  Compare survival curves for different levels of the categorical factor
  log_rank_result <- survdiff(Surv(CorrectedSurvival12, DeathStatus) ~
  ↪  as.factor(projdata[[factor_name]]), data = projdata)
  print(paste("Log-Rank Test for", factor_name))
  print(log_rank_result)

}




surv_obj <- Surv(projdata$CorrectedSurvival12, projdata$DeathStatus)


cox_model <- coxph(
  surv_obj ~ Gender + Age + AgeGroup + Smoker + AbstinenceStatus + HBV + HCV +
    Other + Screening + TumorSize + Diabetes + Alcohol_consumption +
    Thrombosis + Criteria + CancerStages + DifuseCancer + MetastaticCancer +
    CurativeTreatment + Treatment1 + Treatment2 + Treatment3 + Treatment4 +
    Treatment5 + Treatment6,
  data = projdata
)


summary(cox_model)


cox_model <- coxph(
  surv_obj ~ Gender + Age + AgeGroup + Smoker + AbstinenceStatus + HBV + HCV +
     Screening + TumorSize + Diabetes + Alcohol_consumption +
    Thrombosis + Criteria + CancerStages + DifuseCancer + MetastaticCancer +
    CurativeTreatment + Treatment1 + Treatment2 + Treatment3 + Treatment4 +
    Treatment5 + Treatment6,
  data = projdata
)


summary(cox_model)
```

```r
cox_model <- coxph(
  surv_obj ~ Gender + Age + AgeGroup + Smoker + AbstinenceStatus + HBV + HCV +
    Screening + TumorSize + Diabetes + Alcohol_consumption +
    Thrombosis + Criteria + CancerStages + DifuseCancer + MetastaticCancer +
    CurativeTreatment + Treatment1 + Treatment2 + Treatment4 +
    Treatment5 + Treatment6,
  data = projdata
)

summary(cox_model)


cox_model <- coxph(
  surv_obj ~ Gender + Age + AgeGroup + Smoker + AbstinenceStatus + HBV + HCV +
    Screening + TumorSize + Diabetes + Alcohol_consumption +
    Thrombosis + Criteria + CancerStages + DifuseCancer + MetastaticCancer +
    CurativeTreatment + Treatment1 + Treatment2 + Treatment4 +
    Treatment5 ,
  data = projdata
)

summary(cox_model)

cox_model <- coxph(
  surv_obj ~  Age + AgeGroup + Smoker + AbstinenceStatus + HBV + HCV +
    Screening + TumorSize + Diabetes + Alcohol_consumption +
    Thrombosis + Criteria + CancerStages + DifuseCancer + MetastaticCancer +
    CurativeTreatment + Treatment1 + Treatment2 + Treatment4 +
    Treatment5 ,
  data = projdata
)


summary(cox_model)

cox_model <- coxph(
  surv_obj ~  Age + AgeGroup + Smoker + AbstinenceStatus + HBV + HCV +
    Screening + TumorSize + Diabetes + Alcohol_consumption +
    Thrombosis + Criteria + CancerStages + DifuseCancer + MetastaticCancer +
    CurativeTreatment + Treatment1 + Treatment2 + Treatment4  ,
  data = projdata
)


summary(cox_model)

cox_model <- coxph(
  surv_obj ~  Age + AgeGroup + Smoker + AbstinenceStatus + HBV + HCV +
    Screening + TumorSize + Diabetes + Alcohol_consumption +
    Thrombosis + Criteria + CancerStages + DifuseCancer + MetastaticCancer +
    CurativeTreatment + Treatment1 + Treatment2   ,
  data = projdata
)

summary(cox_model)
```

```r
cox_model <- coxph(
  surv_obj ~  Age + Smoker + AbstinenceStatus + HBV + HCV +
    Screening + TumorSize + Diabetes + Alcohol_consumption +
    Thrombosis + Criteria + CancerStages + DifuseCancer + MetastaticCancer +
    CurativeTreatment + Treatment1 + Treatment2   ,
  data = projdata
)


summary(cox_model)

cox_model <- coxph(
  surv_obj ~  Age + Smoker + AbstinenceStatus + HBV + HCV +
    Screening + TumorSize + Diabetes + Alcohol_consumption +
    Thrombosis + Criteria + CancerStages + DifuseCancer + MetastaticCancer +
    CurativeTreatment + Treatment1    ,
  data = projdata
)


summary(cox_model)

cox_model <- coxph(
  surv_obj ~  Age + Smoker   + HBV + HCV +
    Screening + TumorSize + Diabetes + Alcohol_consumption +
    Thrombosis + Criteria + CancerStages + DifuseCancer + MetastaticCancer +
    CurativeTreatment + Treatment1    ,
  data = projdata
)


summary(cox_model)

cox_model <- coxph(
  surv_obj ~  Age + Smoker   + HBV + HCV +
    Screening + TumorSize + Diabetes + Alcohol_consumption +
    Thrombosis + Criteria + CancerStages + DifuseCancer + MetastaticCancer +
    CurativeTreatment + Treatment1    ,
  data = projdata
)


summary(cox_model)

cox_model <- coxph(
  surv_obj ~  Age + Smoker   +  HCV +
    Screening + TumorSize + Diabetes + Alcohol_consumption +
    Thrombosis + Criteria + CancerStages + DifuseCancer + MetastaticCancer +
    CurativeTreatment + Treatment1    ,
  data = projdata
)


summary(cox_model)
```

```
cox_model <- coxph(
  surv_obj ~   Smoker  +  HCV +
    Screening + TumorSize + Diabetes + Alcohol_consumption +
    Thrombosis + Criteria + CancerStages + DifuseCancer + MetastaticCancer +
    CurativeTreatment + Treatment1    ,
  data = projdata
)

summary(cox_model)

cox_model <- coxph(
  surv_obj ~   Smoker  +  HCV +
    Screening + TumorSize + Alcohol_consumption +
    Thrombosis + Criteria + CancerStages + DifuseCancer + MetastaticCancer +
    CurativeTreatment + Treatment1    ,
  data = projdata
)


summary(cox_model)

cox_model <- coxph(
  surv_obj ~   Smoker  +
    Screening + TumorSize + Alcohol_consumption +
    Thrombosis + Criteria + CancerStages + DifuseCancer + MetastaticCancer +
    CurativeTreatment + Treatment1    ,
  data = projdata
)


summary(cox_model)
```

For access to the R codes and dataset files, please visit the following GitHub repository: Lung-Cancer-Analysis