

## Project Proposal

**Leader:** Chenhui Yan (cy2772)

**Team Members:** Yuting Gu (yg3041), Erin Ge (yg3040), Yuechu Hu (yh3822), Kai Tan (kt3026)

**Category:**

Analysis survival data from real-world applications

**Topic:**

Factors Associated with Survival in Patients with Advanced Lung Cancer

**Objective:**

The purpose of this project is to investigate the factors that influence survival time among patients with advanced lung cancer.

**Data Description:**

The dataset is the lung cancer dataset from the R survival package, containing 228 observations and 10 variables. It includes 6 numeric variables ('time', 'age', 'ph.karno', 'pat.karno', 'meal.cal', 'wt.loss'), 1 binary variable 'status', 1 ordinal variable 'ph.eco', and 2 categorical variables ('sex', 'inst'). Among them, the survival information is 'time' and 'status'. There are 67 missing values in this dataset, with 47 missing in the variable 'meal.cal', 14 missing in the variable 'wt.loss', 3 missing in the variable 'pat.karno', and 1 missing in each of the variables 'inst', 'ph.karno', 'pat.karno'. More specifically,

- inst: Institution code.
- time: Survival time in days.
- status: Censoring status 1=censored, 2=dead.
- age: Age in years.
- sex: Male=1, Female=2.
- ph.ecog: ECOG performance score as rated by the physician. 0=asymptomatic, 1=symptomatic but completely ambulator.
- ph.karno: Karnofsky performance score (bad=0-good=100) rated by physician.
- pat.karno: Karnofsky performance score as rated by the patient.

- meal.cal: Calories consumed at meals.
- wt.loss: Weight loss in the last six months (pounds).

### **Statistical Analysis Plan:**

We will first perform descriptive statistics to summarize patient characteristics and explore missing data patterns. Kaplan–Meier curves will be used to estimate and visualize survival probabilities, and log-rank tests will compare survival between groups. Next, univariate Cox proportional hazards models will identify variables associated with survival. Significant and clinically relevant variables will then be included in a multivariable Cox model to estimate adjusted hazard ratios. Given the presence of tied event times in the lung cancer dataset, where multiple patients experience events on the same day, we will compare the Efron and Breslow methods for handling ties in the partial likelihood estimation. The Efron method will be used for final analysis as it provides more accurate estimates, though we will assess whether method choice substantially impacts our results. The proportional hazards assumption will be evaluated using Schoenfeld residuals and log–log survival plots. All analyses will be conducted in R.

### **Timeline:**

By 11/16: Conduct data cleaning and descriptive analysis.

By 11/23: Perform Kaplan–Meier and log-rank analyses.

By 11/30: Build Cox proportional hazards models and check assumptions.

By 12/07: Interpret results and finalize the report.

### **Roles:**

Chenhui Yan (cy2772) and Yuting Gu (yg3041): Build Cox proportional hazards models and check assumptions.

Erin Ge (yg3040): Perform Kaplan–Meier and log-rank analyses.

Yuechu Hu (yh3822): Conduct proposal, data cleaning, and descriptive analysis.

Kai Tan (kt3026): Interpret results and finalize the report.

Currently, we plan for all members to participate in creating the presentation slides and the presentation.

