

**Survival analysis project**

Prostate cancer

Professor: Anna Freni Sterrantino

Students: Thao Pham

Gilles Degue

Rhosane Santos

Shirin Abousalih

Summary

[1. Aim of the project 2](#_Toc1093783031)

[2. Data description 2](#_Toc39623563)

[3. Summary survival statistics 3](#_Toc1047703578)

[4. Survival curves plots and curves comparison 3](#_Toc270276430)

[5. Non-parametric estimation of survival of 1 or more groups 3](#_Toc968796198)

[6. Nonparametric comparison of 2 or more groups 4](#_Toc1539962941)

[7. Semi-parametric Cox regression 5](#_Toc1577598264)

## Aim of the project

The primary goal of this project is to understand the factors influencing the survival rates of prostate cancer patients and to apply statistical techniques to analyze survival times.

Objectives:

Perform data cleaning and preprocessing on prostate cancer datasets. Apply Kaplan-Meier estimators to estimate survival functions. Use Cox proportional hazards models to identify significant predictors of survival. Interpret the results and draw meaningful conclusions about prostate cancer survival.

## Data description

The dataset contains 14,294 rows and 5 columns, Figure 1. The features are described below:

**Grade:** a factor with levels mode (moderately differentiated) and poor (poorly differentiated)

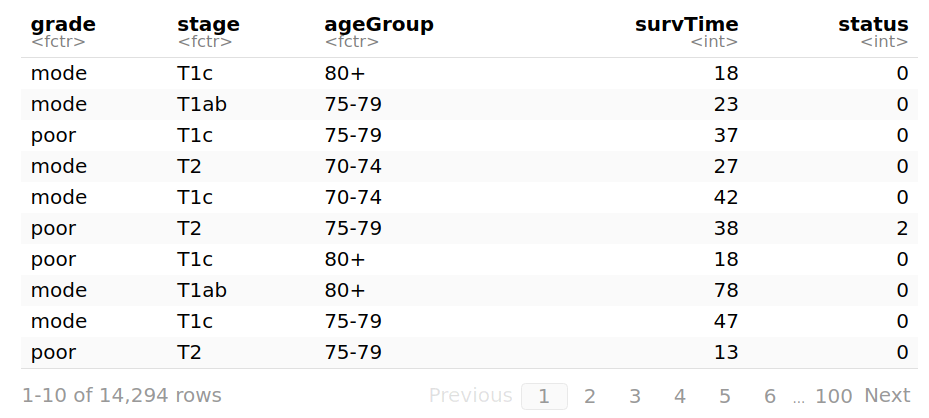
**Stage:** a factor with levels T1ab (Stage T1, clinically diagnoseed), T1c (Stage T1, diagnosed via a PSA test), and T2 (Stage T2)

**AgeGroup:** a factor with levels 66-69 70-74 75-79 80+

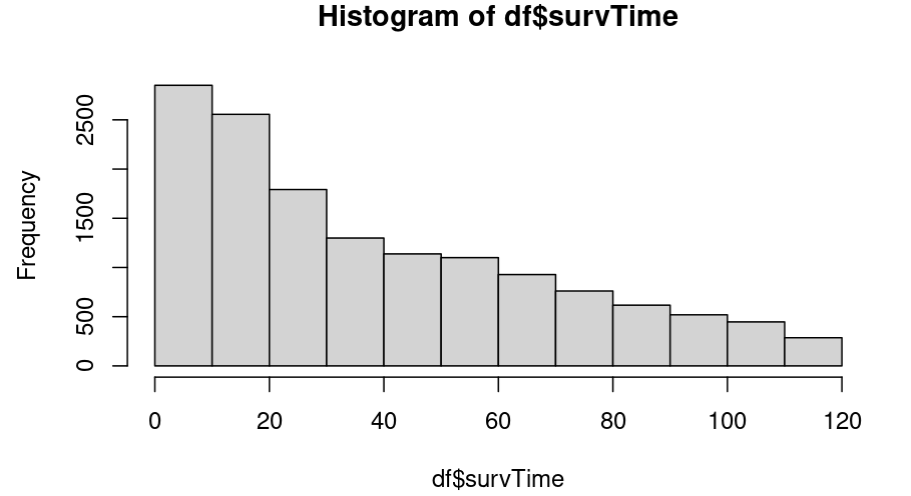
**SurvTime:** time from diagnosis to death or last date known alive

**Status:** a censoring variable, 0, (censored), 1 (death from prostate cancer), and 2 (death from other causes)

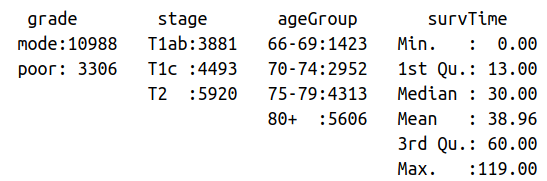
In Figure 2, we can see that the survival time decreases for a higher population. Main descriptive information about the features is shown in Figure 3.



**Figure 1:** First 10 rows of the dataset.



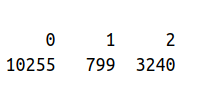
**Figure 2:** Histogram of the survival time.



**Figure 3:** Main statistics of the numerical features.

The quantity of censored data points is shown in Figure 4, where 0 stands for censored data, 1 for prostate related death and 2 for other type of death. But to proceed with the analysis, we removed 2, in other words, we take only into account deaths related to prostate cancer. So, no competing risks.

* 0 (Censored): 10,255 patients (approximately 93%) were censored, meaning they were still alive at the end of the study period or were lost to follow-up before their death.
* 1 (Death due to prostate cancer): 799 patients (approximately 7%) died specifically from prostate cancer.



**Figure 4:** Distribution of censored data in the dataset.

## Summary survival statistics

## Survival curves plots and curves comparison

## Non-parametric estimation of survival of 1 or more groups

Using the Kaplan-Meier estimator, we determine the median survival time which is 32 days, Figure . We see the probability of surviving after the diagnosis in days in Figure . The Nelson-AAlen estimator gives the same results.

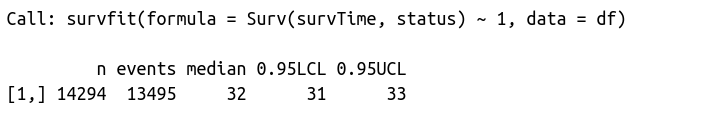


Figure: Median survival time is 32 days.

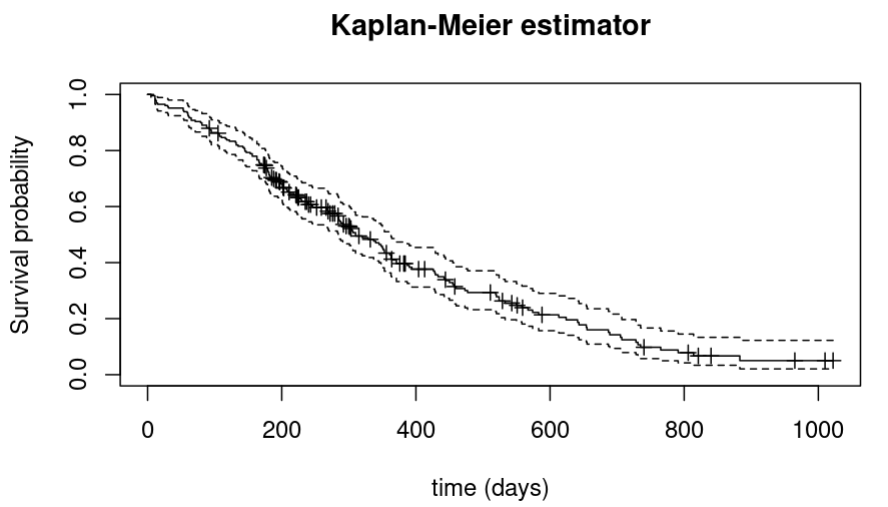


Figure : Survival probability in each time after the diagnosis.

## Nonparametric comparison of 2 or more groups

According to the Figure , we see if there was a difference in survival time according to status through the p-value. As the p-value is <5%, there was a significant difference according to status.

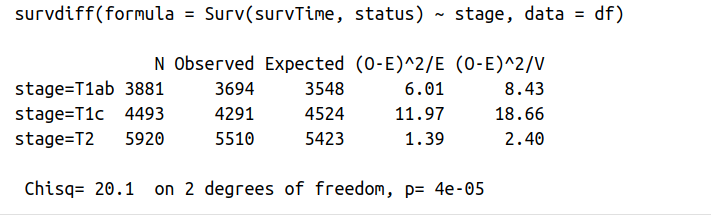
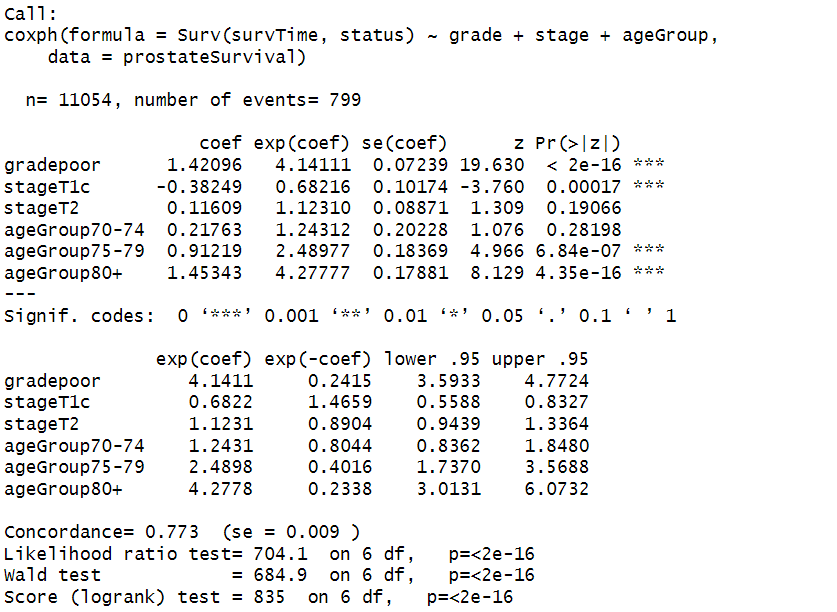


Figure : Log-rank p-value used to determine difference in survival time according to a group.

## Semi-parametric Cox regression

Figure: Summary of the cox model with all variables

The Cox model reveals that poor grade and older age significantly increase the risk of the event, while stage T1c is associated with a lower risk.We can note that the patients with a poor grade have a hazard ratio (HR) of 4.14, indicating they are over four times more likely to experience the event compared to the others. For stage T1c, the HR is 0.68, suggesting a lower risk compared to the reference group, with a significant result, while stage T2 shows a slight, non-significant increase in risk (HR = 1.12). For the AgeGroups, the 75-79 age group has a significant HR of 2.49, and the 80+ group has an HR of 4.28, indicating a substantially higher risk of the event as age increases. Overall, the model highlights

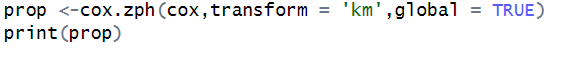
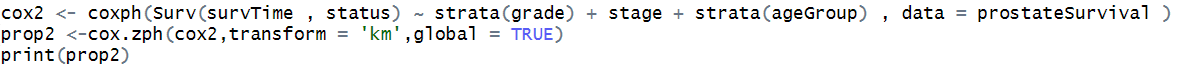
 

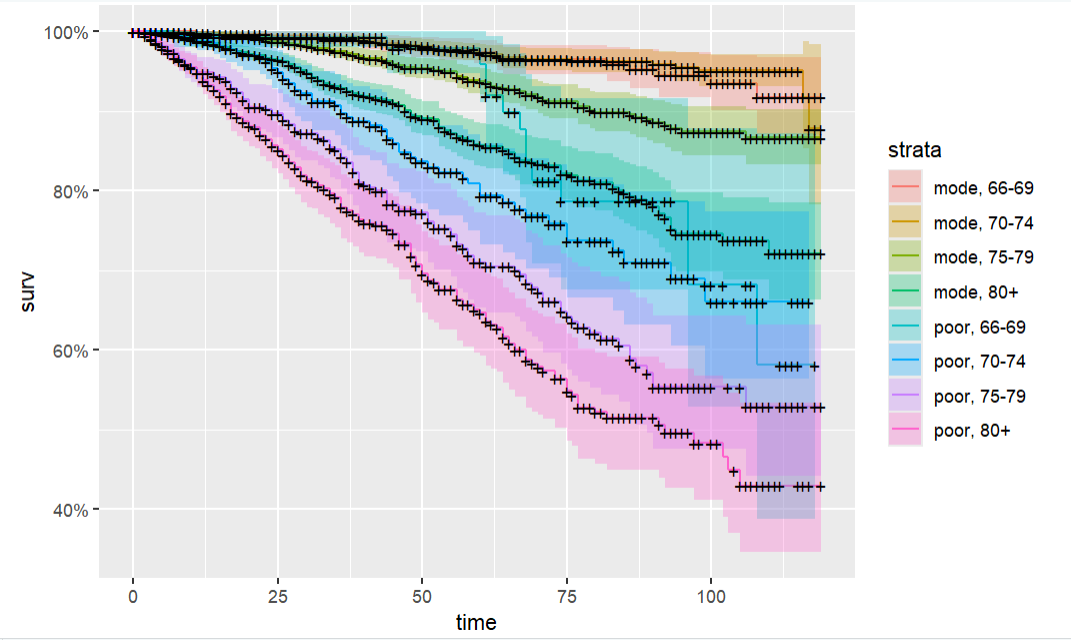
Figure: Verification of the proportional hazard assumption hypothesis

The variables "grade" and "ageGroup" violate the proportionality assumption (p-values of 0.00925 and 0.00104, respectively). The assumption holds for the "stage" variable (p-value of 0.09030). The global test has a p-value of 0.00011 the model violates the proportional hazards assumption.

We also used the stepwise selection process for the Cox proportional hazards model retained all the variables in the model. It seems that removing any of them resulted in a worse fit.

Figure: Verification of the proportional hazard assumption hypothesis

After stratifying the Cox model by grade and age group, we can see a notable improvement in the model. The tests indicate that both the variable "stage" and the global model now respect this assumption (p > 0.05). The effects of covariates on the event risk are constant over time.

Figure: Cox model stratified

The graph shows that survival curves vary significantly according to strata defined by grade and ageGroup. Patients belonging to “poor” grade strata and older age groups (70-74, 75-79, 80+) have lower survival rates, while younger patients and those with “moderate” grade have better chances of survival.