Throat Cancer Analysis

Matthew Murnane

Abstract

Introduction

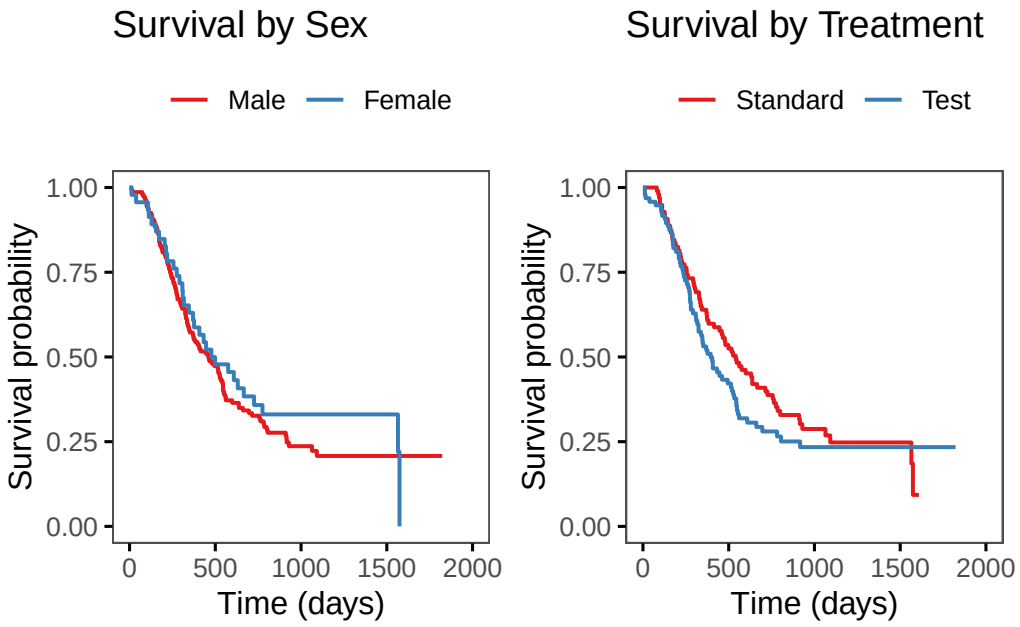
Motivation
About the Data
Exploratory Data Analysis

Table 1: Distributions Status, Sex, Treatment (Tx), and T_Stage

Censored	Dead	Male	Female	Standard	Test	< 2cm	2cm-4cm	> 4cm	massive
53	140	147	46	98	95	9	26	92	66

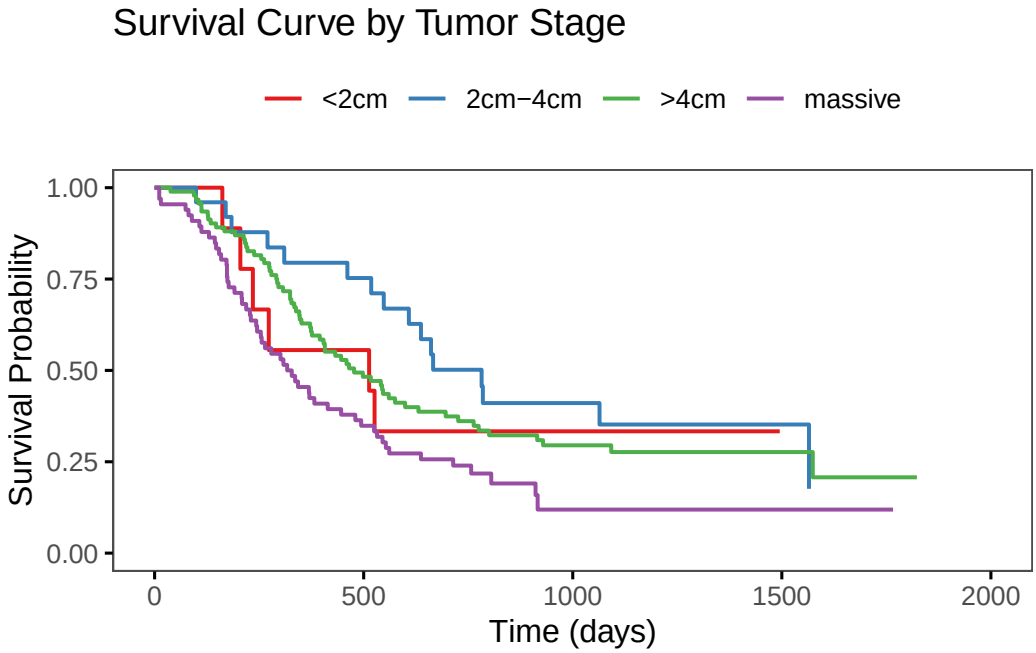
The table above shows distributions for variables of interest: Status, Sex, Tx, and T_Stage. We see that we have 53 censored variables. This would mean that 27.46% of our observations are censored. There is also a disproportionate number of Males to Females in this study. The treatment (Tx) groups are balanced. T_stage is unbalances with the majority of our cases being severe. That is tumor sizes being either greater than 4cm or classified as massive by the study. This a variable we suspect to affect treatment (Tx) effectiveness.

‘‘‘

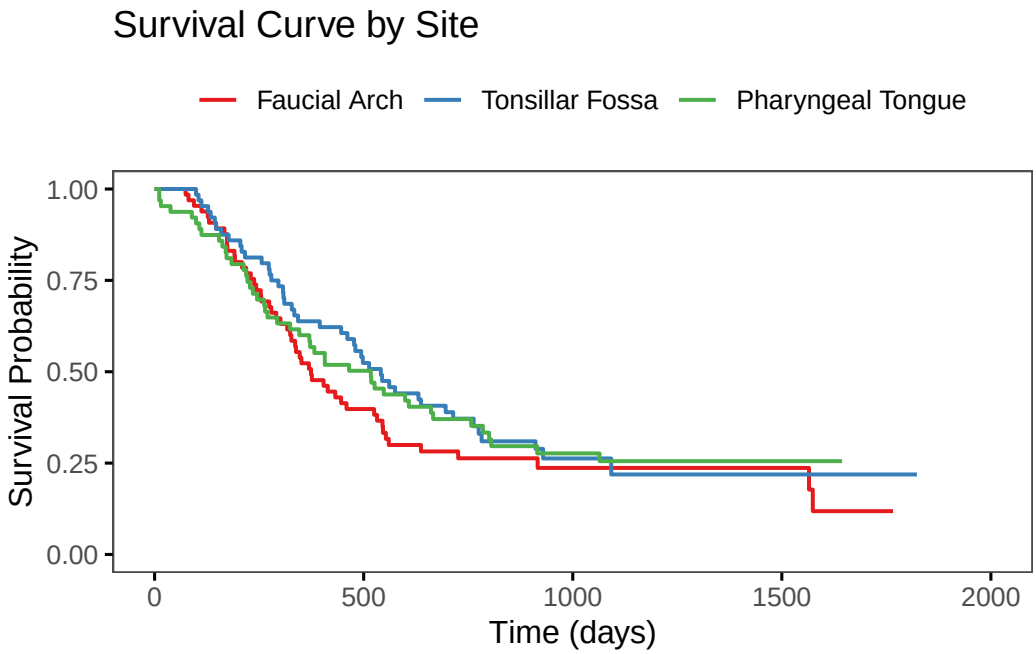


Above we have the Kaplan-Meier survival curves stratified by `Sex` and Treatment (`Tx`). Looking at the plot for sex we see that the curves are pretty much on top of each other until later in the study. The plateau in the female group could be a consequence of the smaller sample size. We are not too worried about the divergence between the curves towards the end of the study but will check if Sex violates Proportional Hazard Assumption when we do Cox Regression.

Looking now at the Survival plot for Treatment we see that in the beginning the curves stay on top of each other but by day 300 they diverge with the standard treatment having a higher survival time. They end up aligning again by day 1000 and the survival time for those in the test group end up higher than in the standard.



Now looking at the KM-curve stratified by Tumor Stages we see a clear difference in survival times among `T_Stage`. We see the most serious case, a tumor classified as "massive" have the steepest drop off in survival time and remains the lowest through out the study. The next lowest is a tumor classified as being greater than 4 centimeters. This variable will be checked for violation of the proportional hazard assumption when we do cox regression.



Above we have the Survival curves for `Site`. They are mostly similar other than around day 400 by the suffers of Facial Arch. All the curves plateau together at .25 around day 1000. It does not look like these variables are significantly different.

Methods

Two treatments were provided in this study and its important to assess if one is better than the other. We will use a Stratified Log Rank Test to asses if the new `test` treatment is better than `standard` by controlling for the variable `T_Stage`. Many variables in the study assess a the severity of a patients diagnosis, we believe the `T_Stage` does the best in consolidating that information. Other variables of that we will stratify will be `Sex` and `Site`.

We are then interested in estimating the hazard rate of a patients cancer journey by taking into account all variables included in the study and will do so using Cox Proportional Hazard Model.

Stratified Log Rank Test

Our null hypotheses is that there is no difference between the hazard functions of treatment groups within each level of `T_Stage`. Our null is that at least one hazard function of treatment differs within a group.

Formally:

$$\mathrm{H}_0: \quad \lambda_1(t|\mathrm{Tx}) = \lambda_2(t|\mathrm{Tx}) \quad \forall t, \mathrm{Tx} \tag{1}$$
$$\mathrm{H}_1: \quad \lambda_1(t|\mathrm{Tx}) = \theta\lambda_2(t|\mathrm{Tx}) \quad \forall t, \mathrm{Tx} \tag{2}$$

The assumption of the Log Rank Test are as follows:

- Censoring is unrelated to a prognosis.
- The survival probabilities are the same for subjects recruited earlier and later in the study
- The events happened at the time specified

Cox Proportional Hazard Model

Cox Proportion Hazard Model takes the form:

$$\lambda(t, \vec{x}_i) = \lambda_0(t)\text{exp}(\vec{x}_i\vec{\beta})$$

Where $\vec{x}_i$ is the vector of covariates for the i^{th} observation and $\vec{\beta}$ are the parameters we will estimate with a partial likelihood function. Understand that $\lambda_0(t)$ is the baseline hazard function. It is the risk of failure at time t when all covariates are zero. $\text{exp}(\vec{x}_i\vec{\beta})$ is the scaling factor. It is a function of covariates.

The Proportional Hazard Assumption of Cox Proportional Hazard Model: Consider two individuals i and j, each with their own vector of covariates.

$$\frac{\lambda(t, \vec{x}_i)}{\lambda(t, \vec{x}_j)} = \frac{\lambda_0(t)\text{exp}(\vec{x}_i\vec{\beta})}{\lambda_0(t)\text{exp}(\vec{x}_j\vec{\beta})} = \text{exp}(\vec{\beta}(\vec{x}_i - \vec{x}_j)) = \theta$$

The hazard ratio of two individuals remains proportional over time.

Results

Stratified Log Rank Test

Our first Log Rank test only considered treatment (`Tx`). It yielded a p-value of .291. We will fail to reject the null that the treatments have differing hazard rates.

Table 2: Stratified Log-Rank Test for Tx

Chi.square	df	p.value
1.11	1	0.291

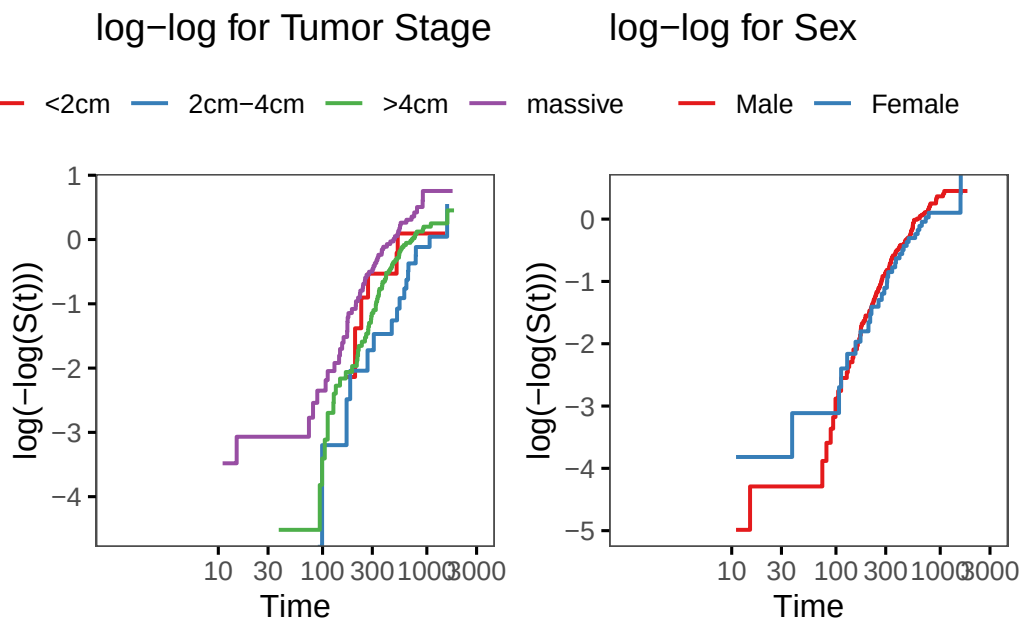
We also ran Log Rank Tests stratified for `T_Stage`, `Sex`, and `Site`. As you can see none of them were found to be significant. Thus we fail to reject the null that the hazard rates between `Tx` within their respective groups differ.

Table 3: Log Rank Results for T_Stage, Sex, and Site

Chi.square	df	p.value	Chi.square	df	p.value	Chi.square	df	p.value
1.39	1	0.239	1.3	1	0.254	1.3	1	0.254

Cox Proportional Hazard Model

Below we check the proportion hazard assumption using log-log survival curves one our three variables of interest. We see in the first graph `T_Stage` is proportional for all factors except tumors < 2cm. The < 2cm red line is very step like. This makes us think that the small size of the group is contributing to it not being proportional to the rest. `Sex`, although it crosses several time is basically on top of each other and considering it's long rank was insignificant I would assume its θ to be 1.



`Faucial Arch` violates the PH assumption. It crosses both levels. Lastly `Tx` crosses at day 100 and seems to diverge around day 300. The divergences is not severe considering the log-log emerges at the end. None of the graphs show a perfect proportional hazard but none of them, except `Site`, are exceptional in their violation. We will run a Cox Proportional Hazard Model with stratified by `Site` with `Tx` ,`Sex`, and `T_Stage` as standard covariates.

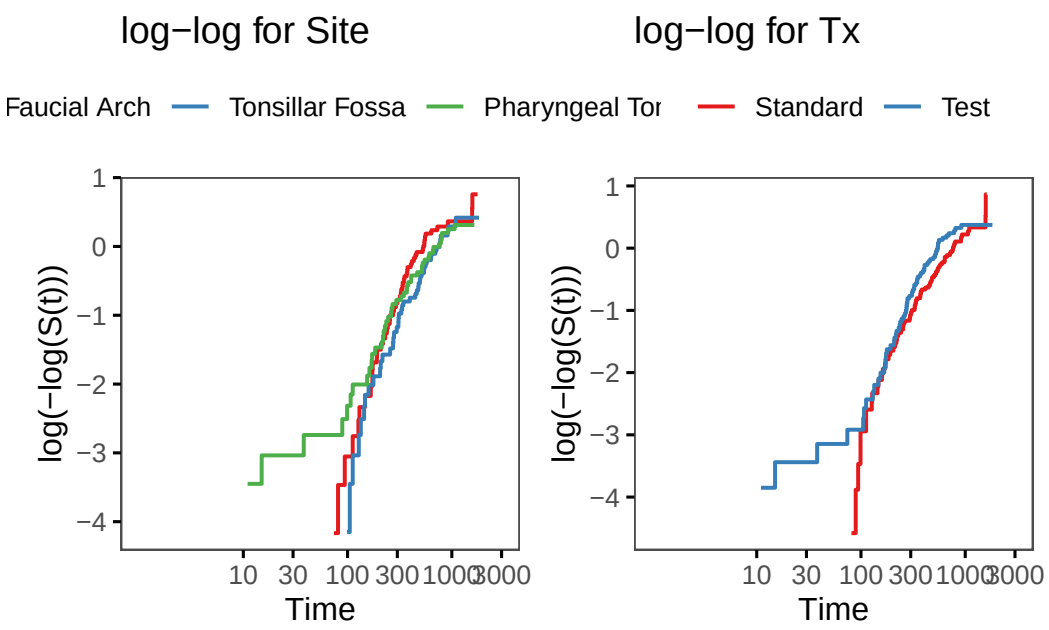


Table 4: Cox Proportional Hazards Model Summary

Variable	Coefficient	Hazard.Ratio	p.value
TxTest	0.172	1.188	0.325
SexFemale	-0.067	0.935	0.745
T_Stage2cm-4cm	-0.294	0.745	0.546
T_Stage>4cm	0.019	1.019	0.965
T_Stagemassive	0.468	1.597	0.286

First we see that none of our coefficients are significant at even the most liberal level of 10%. The model seems to estimate a 19% increase in hazard rate for the `test` treatment. A 6.5% decrease in hazard rate for `Female`. A 25.5% decrease in hazard rate for tumor size between

Conclusion