Lifetime Data Analysis

Practical work

Maria Antònia Colomar and Arnau Garcia

A project presented for the course of
Lifetime Data Analysis 2023/24 (MESIO)

Facultat de Matemàtiques i Estadística
Universitat Politècnica de Catalunya
Date of submission: 14/Jan/2024
Teachers: Guadalupe Gómez and Klaus Langohr

Contents

1 Introduction

Currently, prostate cancer is the most common tumor in men in Spain, with 29.002 new diagnoses in 2023 [1]. Although the survival rate for this tumor is high, in 2020 it caused a total of 5922 deaths in Spain.

These data can be generalized to the rest of the world, as prostate cancer is one of the most common cancers in men in all parts of the world. That is why there are a large number of studies related to this disease (more than 200 thousand studies in *PubMed* [2]).

Hormonal therapies have been widely used with this type of cancer. Hormonal therapies are used to stop the growth of cancer-causing hormones and prevent tumors from proliferating further.

Estrogen is a category of sex hormone responsible for the development and regulation of the female reproductive system and secondary sex characteristics. There are hormonal therapies for the prostate cancer that use Estrogen. According to the *National Cancer Institute*: *Estrogens avoid the bone loss seen with other kinds of hormone therapy, but they increase the risk of cardiovascular side effects, including heart attacks and strokes. Because of these side effects, estrogens are rarely used today as hormone therapy for prostate cancer* [3].

In this practical work we will be analyzing a data set from a study on Prostate Cancer. The data set includes data from 450 patients who have received estrogen treatment. Our goal in this work will be to analyze the efficacy of estrogen treatments on the survival of cancer patients accounting for other variables such as age, cancer stage, or blood pressure. In addition, in the data set we have recorded whether or not the patients have had any cardiovascular disease, which is a factor that is related with the estrogen treatments as we commented.

[1] See `https://seom.org/publicaciones/el-cancer-en-espanyacom` for more information.

[2] See `https://pubmed.ncbi.nlm.nih.gov/?term=prostate+cancer`, the *PubMed* webiste.

[3] See `https://www.cancer.gov/types/prostate/`

2 Descriptive analysis

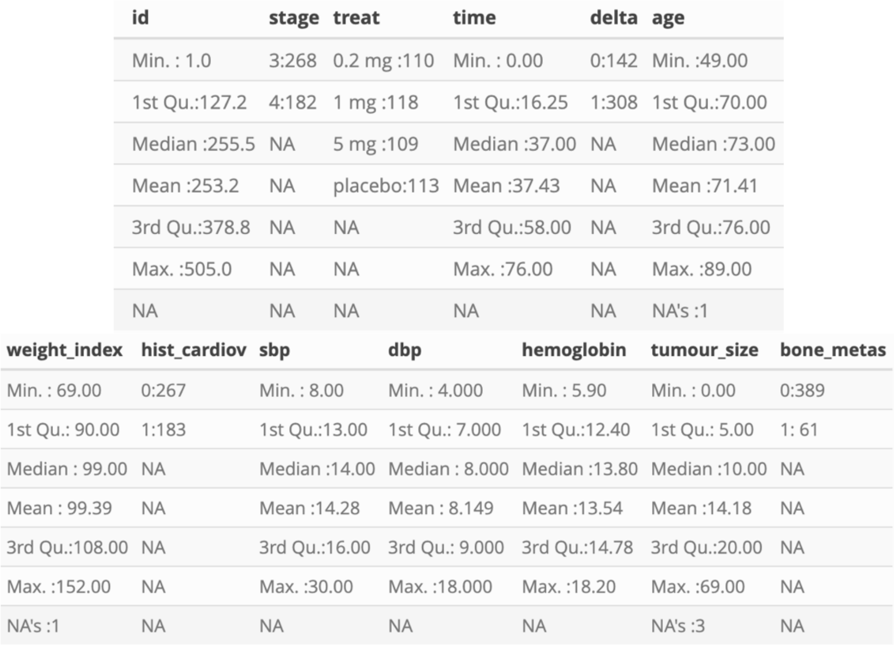
We have a data set with 450 observations corresponding to 450 different patients with prostate cancer. We have a total of 13 variables. In this section our objective is to study the data set and its variables, understand the different items in our data and make the necessary comments before the analysis. The reader can see a table with a summary statistics of all the variables in the Figure 1. In this table we have included the minimum, first quartile, median, mean, third quartile, maximum, and number of missing values (appears as NA) for the numerical variables. And, for the categorical variables, we have included the frequencies for each category, and the number of missing values.

Let us comment separately the variables of our data set. We start with the numerical variables.

- **Id**: The patient identifier. There is exactly one patient id for each patient.
- **Time**: Survival time from study start in months. The main comment on this variable is that we have observations that are 0. The reader can easily observe this in the table, since the minimum recorded is 0. Having null survival times is a problem for our analysis, and it is also an unrealistic data, it does not make sense that there are patients who die exactly at the time of starting the study. So, what we have decided to do with these data is to add 0.5 to them.
- **Age**: Age of the patient. The youngest patient is 49 years old, and the oldest is 89 years old, while the mean age of the patients in the study is 71.41 years old. This fits with the age at which males typically present with this type of cancer.
- **Weight index**: Weight index for each patient, which is computed using the formula: weight$(kg) -$ height$(cm) + 200$. Obviously the weight index is an indicator of the patient's condition. Values close to the minimum and maximum will be indicators of poor patient health.
- **SBP**: Systolic blood pressure (divided by 10).
- **DBP**: Diastolic blood pressure (divided by 10).
- **Hemoglobin**: Hemoglobin level in g/100ml.
- **Tumour size**: Tumour size in cm^2. Clearly, tumor size is directly related to disease stage.

Now, we analyze the categorical variables.

- **Stage**: Cancer stage. The reader can observe in the Table in Figure 1 that the patients on the study have a cancer stage 3 or 4, these are the two worst stages of cancer. We have 268 patients with Cancer Stage 3, and 182 patients with Cancer Stage 4. We can intuit that in this work this variable will not have a very great impact, or at least not as great an impact as it would have if the four stages of cancer were represented in our sample.
- **Treat**: Estrogen treatment: Placebo, 0.2 mg, 1 mg, or 5 mg of estrogen. This is the main variable of the data set, because our objective in this work is analyze which is the most effective treatment. We have four different estrogen treatments (including placebo), and as we can see in Figure 1 our data is balanced, in terms that we have approximately the same number of patients receiving each treatment. In addition, we have no missing values. These two facts are positive for the further analysis.
- **Delta:** Death indicator, where 1 means "yes" and 0 "no". We have 308 events versus 142 censored times. It seems a good ratio for our purposes.
- **Hist. Cardiov.**: History of cardiovascular disease where 1 means "yes", it is, that the patient has had cardiovascular disease. And 0 means "no", it is, the patient has not suffered from cardiovascular disease. The 59.3% of patients have not suffered from cardiovascular disease.



id	stage	treat	time	delta	age
Min. : 1.0	3:268	0.2 mg :110	Min. : 0.00	0:142	Min. :49.00
1st Qu.:127.2	4:182	1 mg :118	1st Qu.:16.25	1:308	1st Qu.:70.00
Median :255.5	NA	5 mg :109	Median :37.00	NA	Median :73.00
Mean :253.2	NA	placebo:113	Mean :37.43	NA	Mean :71.41
3rd Qu.:378.8	NA	NA	3rd Qu.:58.00	NA	3rd Qu.:76.00
Max. :505.0	NA	NA	Max. :76.00	NA	Max. :89.00
NA	NA	NA	NA	NA	NA's :1

weight_index	hist_cardiov	sbp	dbp	hemoglobin	tumour_size	bone_metas
Min. : 69.00	0:267	Min. : 8.00	Min. : 4.000	Min. : 5.90	Min. : 0.00	0:389
1st Qu.: 90.00	1:183	1st Qu.:13.00	1st Qu.: 7.000	1st Qu.:12.40	1st Qu.: 5.00	1: 61
Median : 99.00	NA	Median :14.00	Median : 8.000	Median :13.80	Median :10.00	NA
Mean : 99.39	NA	Mean :14.28	Mean : 8.149	Mean :13.54	Mean :14.18	NA
3rd Qu.:108.00	NA	3rd Qu.:16.00	3rd Qu.: 9.000	3rd Qu.:14.78	3rd Qu.:20.00	NA
Max. :152.00	NA	Max. :30.00	Max. :18.000	Max. :18.20	Max. :69.00	NA
NA's :1	NA	NA	NA	NA	NA's :3	NA

Figure 1: Summary table with all the variables of the data set.

- **Bone Metas.**: Bone metastasis, where 1 means "yes", the patient has bone metastasis, and 0 means "no", the patient has not bone metastasis. Only the 13.56% of patients has bone metastasis.

3 Nonparametric analysis

3.1 Estimation of the survival function

In this section we use the Kaplan-Meier estimator for estimate the survival function. Our main interest is to estimate the survival function according to the estrogen treatments, to see how the survival time behaves with respect to the different treatments. This will give us a first idea of which treatments work best. Notwithstanding, we are also interested in study how the survival times behaves with respect other variables, because this will give us an intuition as to which variables most affect our study and in what way they do so. Thus, we have decided to estimate the survival functions according the treatment, the history of cardiovascular disease, the bone metastasis and the cancer stage.

The survival functions according to the treatment are exposed in Figure 2. In the first 20 months the four survival functions are similar, from month 20 (approximately) onwards we can observe differences between the survival functions. Looking at the graphic it is clear that, from the month 20, the treatment with larger survival time is the 1 mg (green line on the graphic) treatment, followed by the treatment of 5 mg (dark blue line). It is more difficult to see weather, on average, the survival times of the 0.2 treatment (red line) are longer than those of the placebo (cyan line), because both survival functions are crossing multiple times and seems to be very near. Notwithstanding, after the analysis of the survival function according to the treatment the idea is clear: the treatment with 1 mg of estrogen seems to be the most effective, followed by the 5 mg, and then the 0.2 mg and the placebo. In the following sections of the work, we will try to demonstrate more rigorously, with the concepts learned during the course, that the above hypothesis holds true.

It is also important to briefly analyze the other survival functions according other variables. These survivals functions are exposed in the Appendix (see Figure 7). In the survival functions estimation according the history of cardiovascular disease we can see that survival times are, on average, longer for those patients who had not previously had a cardiovascular disease. We can also see in the graphic with the survival functions according to the bone metastasis variable that the survival times are, on average, longer for those patients with no bone metastasis. Finally in the graphic that corresponds to the survival functions according to the cancer stage, we can observe that the survival times are, on average, longer for those patients in cancer stage 3.

The behavior of the survival times with respect to these variables is really important and we will have to take it into account throughout the work in order to correctly measure the efficacy of the different treatments taking into account the other variables.

3.2 Estimation of the median survival time

We do an estimation of the median survival time for the survival function estimations according to the treatments. The resulting estimation median survival is:

- Treatment with 0.2 mg of estrogen: 31.5 months.
- Treatment with 1 mg of estrogen: 49.5 months.
- Treatment with 5 mg of estrogen: 36 months.
- Placebo: 36 months.

Thus, these estimation of the median survival times seems to support that the more efficient treatment is the one with 1 mg with estrogen. This is the treatment with the largest median survival time, 49.5 months, which is 13.5 months (more than a year) larger than the second highest median survival time estimation. On the other hand, these estimations put the placebo treatment at the same level than the 5 mg treatment. And if one observe the survival functions in Figure 2, it seems that the 5 mg treatment is a little bit better. Notwithstanding, the difference are rather small and we need more

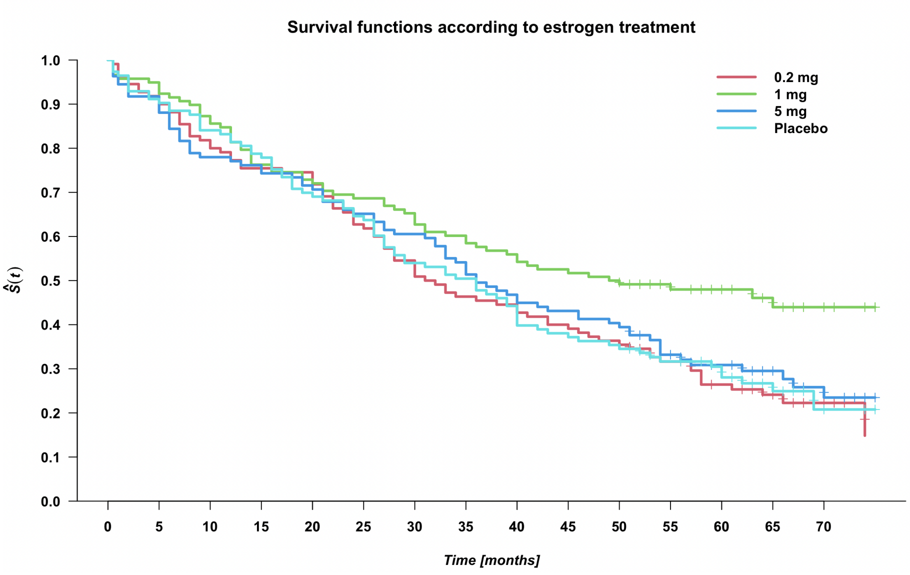


Figure 2: Survival functions according to the treatments.

powerful tools for testing the difference between survival functions, and this is exactly what we will do in the next section.

3.3 Comparison of survival functions

In this section we will be checking if there are significant diffirences between the survival functions according the treatments. Let $S_0(t), S_1(t), S_2(t), S_3(t)$ be the survival functions according to the treatment with 0.2 mg of estrogen, 1 mg, 5 mg and placebo, respectively. Our goal in this section is to conduct the test

$$\begin{aligned} H_0: &\quad S_0(t) = S_1(t) = S_2(t) = S_3(t), \quad \text{for all} \quad t \leq \tau. \\ H_1: &\quad \text{Some} \quad S_i(t) \neq S_j(t), \quad \text{for some} \quad t \leq \tau. \end{aligned}$$

To carry out the previous test we will use the Fleming-Harrington family of tests, as they are a very flexible family of tests from which we can obtain other tests. The Fleming-Harrington family of tests is a weighted family of tests, with weights:

$$W(t_i) = \hat{S}(t_{i-1})^{\rho}(1 - \hat{S}(t_{i-1}))^{\lambda}, \quad \rho, \lambda \geq 0.$$

Thus, different choices of ρ, λ leads to different tests. We have decided to conduct the test with: $\rho = \lambda = 0$, which leads to the classical Logrank Test. Also with, $\rho = 1/2, \lambda = 0$, which leads to a test used for detect early differences. With $\rho = 0, \lambda = 1/2$, which leads to a test used for detect late differences. And finally, $\rho = \lambda = 1/2$.

Let us fix a significance level of $\alpha = 0.05$. The statistic obtained is distributed, under H_0, as a chi square with $4 - 1 = 3$ degrees of freedom. Thus, we reject the null hypothesis if the statistic is large enough, for the significance level fixed if is larger than $\chi^2_3(1-\alpha) = 7.82$. The results obtained for the different choices of ρ, λ are exposed in Table 1.

Test	χ^2_3	Rejection region	p-value
$FH(\rho = 0, \lambda = 0)$	10.1	$10.1 \geq 7.82$	0.0177
$FH(\rho = 1/2, \lambda = 0)$	8	$8 \geq 7.82$	0.0458
$FH(\rho = 0, \lambda = 1/2)$	13.1	$13.1 \geq 7.82$	0.00453
$FH(\rho = 1/2, \lambda = 1/2)$	11.7	$11.7 \geq 7.82$	0.00849

Table 1: Table with the different tests of the Fleming-Harrington family conducted.

We can observe that for the all the different choices of ρ, λ we can reject the null hypothesis and conclude that some $S_i(t) \neq S_j(t)$ for some $t \leq \tau$. Looking at the graphic of the survival functions according to treatment in Figure 2 we see that the results obtained in Table 1 make sense. The survival functions are similar at early times, thus when we take $\rho = 1/2, \lambda = 0$, although we finally reject the null hypothesis, we obtain a higher p-value (0.0458), and we can observe that the statistic obtained is within the rejection region by a small amount ($8 \geq 7$). On the other hand, we can see in the graphic that the survival functions present late differences, then, when we choose $\rho = 0, \lambda = 1/2$ we obtain a smaller p-value (0.00453).

Summarizing, we can conclude that there are significant differences between the survival functions, which supports our hypothesis that some estrogen treatments are more effective than others.

4 Fit of a parametric survival model

4.1 Selecting the appropriate model

In this section we will be fitting and analyzing a parametric survival model. The parametric survival models are log-linear models, this is we have a model with the following form:

$$\log T = \mu + \beta^t Z + \sigma W.$$

Where $\mu \in \mathbb{R}$ is the intercept, $\beta^t = (\beta_1, \cdots, \beta_p)$ is the vector of regression coefficients, $\sigma > 0$, W is the error term distribution and $Z = (Z_1, \cdots, Z_p)$ vector of fixed, i.e., not time-dependent, covariates. The most common parametric choices for T are the Weibull, log-logistic, and lognormal distribution.

The first step is to to check which is the most suitable distribution for our data. In this initial part we do not take into account the variables that we should or should not add to our model. In this part we are only interested in the survival times and events that we have in our data.

We have exposed the cumulative hazards plots in the Figure 3. It seems that the distribution that fits better with the data is the Weibull distribution. Thus, we select this distribution for fit the parametric model. We have also fitted the three different type of parametric models and compared the probability plots. These probability plots support that the best choice is the Weibull distribution. The probability plots will be added to the Appendix (see Figure 8, Figure 9 and Figure 10).

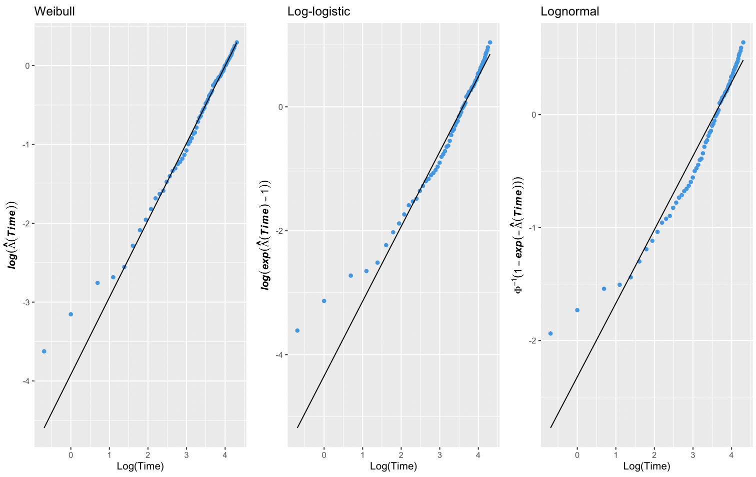


Figure 3: Cumulative hazards plots for, from left to right, the Weibull, the log-logistic and the lognormal distribution.

The following step is to select the best model possible taking into account the different variables available. For this purpose we have followed the next criteria: add the variables that we believe to be essential under medical criteria. And on the other hand, take into account the AIC of each model to select the best one. The variables that, we believe, are the most important under medical criteria after what we have found in the literature and on the internet are: treatment, age, tumour size, history of cardiovascular disease, bone metastasis and the weight index. In addition, of all the models we have adjusted, this is the one with the best AIC (the lowest). The model obtained is exposed in Table 2. It is also important to note that we have tried to add interactions between variables to our model, but the result has not been positive in terms of AIC and significance of the variables. Thus, we chose to work with a model without interactions.

	$\hat{\beta}$	Std. Error	$AF = \exp(-\hat{\beta}_j)$	$HR = \exp(-\hat{\beta}_j/\hat{\sigma})$	p-value
Intercept ($\hat{\mu}$)	4.7	0.81			$6.1 \cdot 10^{-9}$
Treatment 1 mg ($\hat{\beta}_1$)	0.43	0.17	0.6479863	0.6408723	0.0094
Treatment 5 mg ($\hat{\beta}_2$)	0.13	0.16	0.8770671	0.8741449	0.40473
Treatment placebo ($\hat{\beta}_3$)	-0.033	0.15	1.0335489	1.0344170	0.83
Age ($\hat{\beta}_4$)	-0.022	0.0089	1.0218836	1.0224466	0.015
Tumour size ($\hat{\beta}_5$)	-0.017	0.00455	1.0175387	1.0179890	0.00013
Bone metastasis 1 ($\hat{\beta}_6$)	-0.41	0.16	1.5055631	1.5213183	0.01215
History cardiov. 1 ($\hat{\beta}_7$)	-0.5	0.12	1.6481275	1.6692126	$1.6 \cdot 10^{-5}$
Weight index ($\hat{\beta}_8$)	0.012	0.00475	0.9878513	0.9875442	0.01004
Log(scale) ($\log(\hat{\sigma})$)	-0.025	0.05			0.61633

Table 2: Table with the information about the parametric model fitted.

With the model fitted we check graphically whether the Weibull distribution assumption holds. In the Figure 4 we can observe the residual plot obtained. Although the theoretical curve is outside the confidence bands in some parts, the deviation we have is very small. So, using this and what we have seen in the probability plots and the cumulative hazard plots, we believe this is an appropriate model.

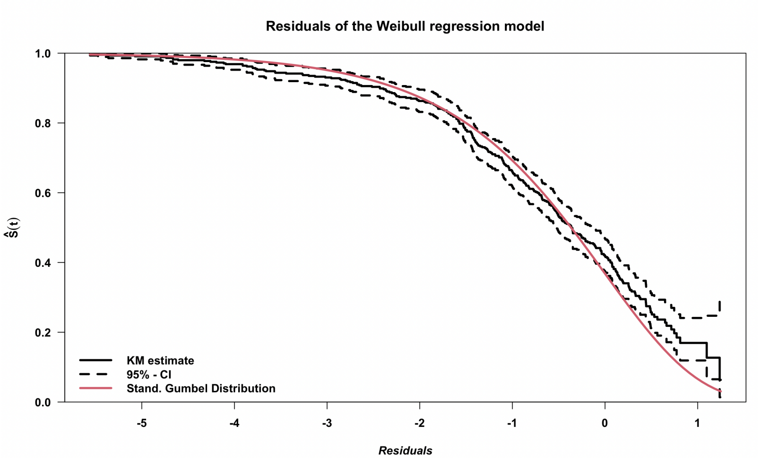


Figure 4: Residual plot to check for the Weibull distribution.

4.2 Interpretation of the model fit

Now, it is time to interpret the model fit, we will be analyzing the results presented in Table 2. According to the model fit, the risk of death is larger

- in the case of those patients receiving Placebo compared to persons who are receiving the treatment with 0.2 mg of estrogen ($\hat{\beta}_3 = -0.33 < 0$).
- the larger the age of the patient ($\hat{\beta}_4 = -0.022 < 0$).
- the larger the tumour size ($\hat{\beta}_5 = -0.017 < 0$).
- in the case of patients who have suffered from cardiovascular disease compared to those who have not ($\hat{\beta}_7 = -0.5 < 0$).

In addition, according to the model, the risk of death is lower in the case of those patients receiving the 1 mg estrogen treatment, and the 5 mg compared to individuals who are receiving the treatment with 0.2 mg of estrogen ($\hat{\beta}_1 = 0.43 > 0, \hat{\beta}_2 = 0.13 > 0$). And, the larger the weight index ($\hat{\beta}_8 = 0.012 > 0$).

We can observe that all the variables added to the Weibull parametric model are significant at a level 0.05. The only variables which are not significant at that level are the Treatment of 5 mg of estrogen and the placebo, but the category Treatment of 1 mg of estrogen is significant, and then the use of this variable is justified. In addition, the treatment variable is the key variable, the one which we are interested. Thus, add this variable to the model is mandatory for our purposes.

4.3 Interpretation of the model parameters

In this section we will be interpreting the model parameters in terms of the relative hazards and the accelerating factor. The reader can see these quantities exposed in Table 2. We study in deep the hazard ratios associated with the treatment variables. We have obtained $\exp{(-\hat{\beta}_1/\hat{\sigma})} = 0.6408723, \exp{(-\hat{\beta}_2/\hat{\sigma}))} = 0.8741449, \exp{(-\hat{\beta}_3/\hat{\sigma}))} = 1.0344170$. Thus, according to the values of the hazard ratios obtained, the instantaneous risk of death of those patients that have received the treatment with 1 mg of estrogen, the treatment with 5 mg of estrogen and placebo, respectively, is 0.64, 0.87 and 1.03 times larger than the one of individuals receiving the treatment with 0.2 mg of estrogen. Then, the instantaneous risk is lower in the case of treatments of 1 mg and 5 mg ($0.64 < 1, 0.87 < 1$) and higher in the case of placebo ($1.03 > 1$) compared with the treatment of 0.2 mg.

Now, we interpret the acceleration factors with respect the categories of the treatment variable. We have obtained $\exp{(-\hat{\beta}_1)} = 0.6479863, \exp{(-\hat{\beta}_2)} = 0.8770671, \exp{(-\hat{\beta}_3)} = 1.0335489$. According to the values of the acceleration factors obtained, the median times until death (or any other quantile) of persons under treatment of 0.2 mg of estrogen is 0.65, 0.88 and 1.03 times larger than the median times of, respectively, the patients receiving the treatment of 1 mg, 5 mg and placebo.

Thus, the conclusion of this interpretations is that the treatments of 1 mg and 5 mg of estrogen are more effective than the treatment of 0.2 mg, since the hazard ratios are $0.64 < 1$ and $0.87 < 1$ respectively. Moreover, the 1 mg treatment seems to be more effective than the 5 mg, since $0.64 < 0.87$. On the other hand, the placebo treatment seems to be the less efficient. Similar conclusions are drawn from the acceleration factors obtained. It is important to mention that all these conclusions are for patients with the same age, tumour size, bone metastasis, history of cardiovascular disease and weight index.

Regarding the other hazard ratios computed and exposed in Table 2, we highlight that the instantaneous risk of death for those patients who have suffered a cardiovascular disease is 1.67 times larger than the one for those patients who have not suffered this type of disease (this means a 67% higher instantaneous risk). In addition, the instantaneous risk for those patients with a bone metastasis is 1.52 times larger than the one for patients without bone metastasis (it is, 52% higher). The previous interpretation for the hazard ratios corresponding with the categorical variables history of cardiovascular disease and bone metastasis, demonstrates that these are important variables in the study that actually affect patient survival and therefore need to be taken into account. The interpretation for the acceleration factors associated with these variables (quantities exposed in Table 2) leads to the same conclusion.

5 Fit of a semi-parametric model

5.1 Cox models or proportional hazards models

The validity of the Cox model is based on the hypothesis that the hazards of two individuals with different covariates are proportional. Violation of the PH assumption may lead to biased effect estimates in Cox regression analysis. In this case we decided to check via a visual assessment of KM curves, we saw on figure 2 that the survivals do not intersect clearly, so we could conclude it is not a sign of the assumption being violated. We will conclude if the model has a biased effect later on in the residuals study.

After fitting the Cox model we obtain the table 3 for the variables involved. We have decided to include the same variables as in the parametric model, as we think it could be more interesting to work with the same selection.

Variable	$\hat{\beta}$	$\exp(\hat{\beta})$	$\exp(-\hat{\beta})$	$s.e.(\hat{\beta})$	p-value
Treat 1 mg ($\hat{\beta}_1$)	-0.453791	0.635215	1.5743	0.170835	0.007900
Treat 5 mg ($\hat{\beta}_2$)	-0.139868	0.869473	1.1501	0.161878	0.387568
Treat placebo ($\hat{\beta}_3$)	0.031111	1.031600	0.9694	0.156562	0.842485
Age ($\hat{\beta}_4$)	0.022407	1.022660	0.9778	0.009119	0.013998
Tumour size ($\hat{\beta}_5$)	0.017691	1.017849	0.9825	0.004631	0.000133
Bone metastasis 1 ($\hat{\beta}_6$)	0.425824	1.530851	0.6532	0.167478	0.011004
History cardiov. 1 ($\hat{\beta}_7$)	0.514954	1.673562	0.5975	0.118069	$1.29{\cdot}10^5$
Weight index ($\hat{\beta}_8$)	-0.012773	0.987308	1.0129	0.004876	0.008804

Table 3: Table with the information about the Cox model fitted

5.2 Interpretation of the model fit

After fitting the model, in this section we will interpret the summary of the Cox model shown in 3. First, we can see that tumor size is very statistically significant, with a positive value coefficient, meaning that the larger the value, the larger the risk. Also weight index seems to be a very statistically significant variable, but with a negative coefficient which means the risk and the value of the variable are inversely proportional. In the same group of significant variables we can notice history of cardiovascular disease with a very small p-value and a positive (and large) coefficient. We find other variables that may be significant like treatment 1mg, with a large negative value, bone metastasis 1 and age with positive values. This gives us an idea on the behavior value of instantaneous risk of death given a variable.

5.3 Interpretation of the model parameters

In this section we interpret the model parameters in terms of the relative hazards. In the Table 3 the reader can observe all the relative hazards correspondent to each variable. According to the results obtained, the instantaneous risk of death,

- is 57% larger for those patients receiving the treatment with 0.2 mg compared with patients under the treatment with 1 mg of estrogen.
- patients receiving the treatment with 0.2 mg of estrogen have a 15% higher instantaneous risk than patients under the treatment with 5 mg of estrogen.

- for those patients receiving placebo the instantaneous risk of death is 3% higher than for those individuals under the treatment with 0.2 mg of estrogen.
- increases with age. The instantaneous risk increase every 5 years by a factor of $\exp(0.022 \cdot 5) = 1.12$.
- increases with tumour size. For every square centimeter of tumour size the instantaneous risk increase by a factor of 1.018.
- is 53% larger for those patients who have a bone metastasis compared with those without metastasis.
- is 67% larger for those patients who have suffered a cardiovascular disease compared to patients who have not suffered from it.
- decreases with the weight index. For 20 units of the weight index the instantaneous risk of death reduces by a factor $\exp(-0.013 \cdot 20) = 0.77$.

Now, it is interesting to compare the hazard ratios between treatments not only with respect the baseline, but with respect all the possibilities. This can be done easily with the Cox model, for instance, if we want to compute the hazard ratio between the treatment of 1 mg and the treatment of 5 mg of estrogen, although the baseline is the treatment with 0.2 mg, we can do the following:

$$HR(Z_1 = 1; Z_2 = 1) = \frac{\lambda(t|Z_1 = 1)}{\lambda(t|Z_2 = 1)} = \frac{\frac{\lambda(t|Z_1=1)}{\lambda_0(t)}}{\frac{\lambda(t|Z_2=1)}{\lambda_0(t)}} = \frac{HR(Z_1 = 1)}{HR(Z_2 = 1)} = \frac{0.635}{0.869} = 0.731.$$

Where $Z_1 = \mathbb{1}_{\{\text{treat 1 mg}\}}, Z_1 = \mathbb{1}_{\{\text{treat 5 mg}\}}$, $HR(Z_1 = 1; Z_2 = 1)$ denotes the hazard ratio of the patients under treatment of 1 mg with respect patients under 5 mg treatment, and $HR(Z_1 = 1), HR(Z_2 = 1)$ denotes the hazard ratio of the patients under the 1 mg, and 5 mg, respectively, with respect the baseline (notice that these last two quantities are both exposed in Table 3). Thus, looking at the inverse of the hazard ratio computed, because is more convenient, we have that the instantaneous risk of death is 37% ($1/0.731 = 1.3685$) higher for those patients under the 5 mg of estrogen treatment compared to patients receiving the treatment with 1 mg.

Using the same, we can compute the relative hazards between all the possible pairs of treatments. The comparison between the 1 mg, 5 mg and the placebo with respect the baseline (the 0.2 mg of estrogen treatment) is already done. Also, the comparison between the 1 and 5 mg treatments. Then, we have yet to compare the 1 mg treatment with the placebo, in this case the instantaneous risk is 62% ($1.0316/0.635 = 1.624$) higher for those patients receiving placebo compared to patients under the 1 mg treatment. And finally, we compare the placebo with the 5 mg treatment, obtaining that the instantaneous risk is 19% ($1.0316/0.869 = 1.187$) higher for the patients receiving placebo compared with the patients receiving 5 mg of estrogen. Thus, in conclusion, the treatment of 1 mg of estrogen is the best, in the sense that it has the lowest instantaneous risk of death compared with the others. Following, we have the 5 mg treatment, then the 0.2 mg and finally the placebo. However, the difference between the placebo and the 0.2 mg treatment is smaller than the difference between the other treatments. Thus, the 0.2 mg treatment does not seem to be much more effective than the placebo.

5.4 Analysis of the residuals

In this section we will analyse the residuals of the model in order to extract some more information about the proportional hazards assumption and influential observations. The proportional hazards assumption can be checked with the output returned by the R function *cox.zph*, shown in the Table 4 which analyses the correlation of the Schoenfeld residuals and survival time. Under the proportional

Variable	χ^2_{df}	df	p-value
treat	3.98	3	0.264
age	5.18	1	0.023
tumour_size	0.0152	1	0.902
bone_metas	0.113	1	0.737
hist_cardiov	0.486e	1	0.486
weight_index	0.000371	1	0.985
GLOBAL	10.7	8	0.217

Table 4: Table with the variable, a score test of for addition of the time-dependent term, the degrees of freedom, and the two-sided p-value.

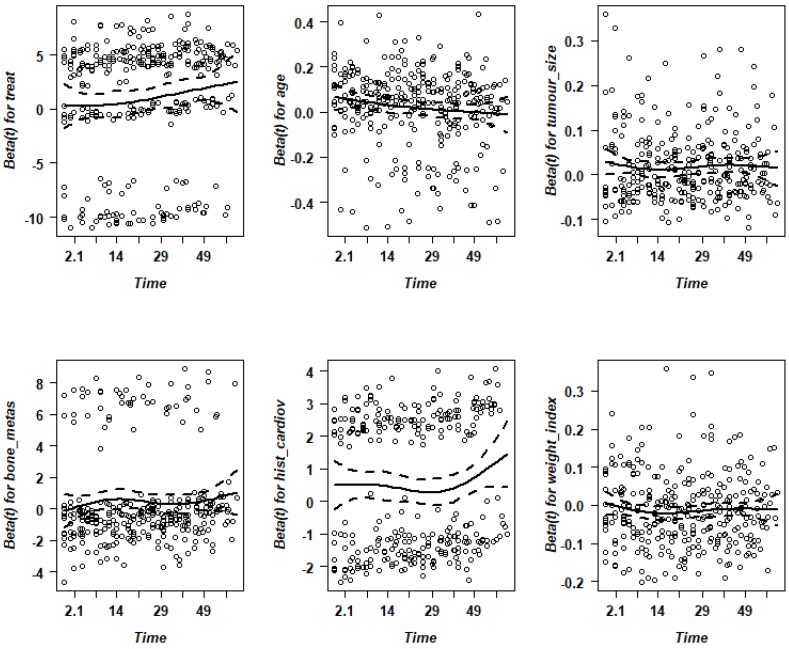


Figure 5: Schoenfeld residuals.

hazards assumption, this correlation is expected to be close to 0 Given that the variable age presents a small value, we cannot assume the proportional hazards (PH) assumptions hold for this variable. We can also see this in the Figure 5 where we see that the line for the estimators for the variables age and history of cardiovascular disease is not straight, proportionality for these variables is in doubt.

We can also look into the dfbeta residuals in Figure 6 which will show if there is evidence of influential observations. We can see a very influential observation in the variable tumor size and history of cardiovascular disease which is the same patient with ID 229 that is a patient in the best treatment of 1mg and with no history of cardiovascular disease with a time survival nearby the mean.

Another influential observation can be seen in bone metastasis that is the patient with ID 166 that presents a large survival time (54, almost third quantile) but 69 years old and with history of cardiovascular disease.

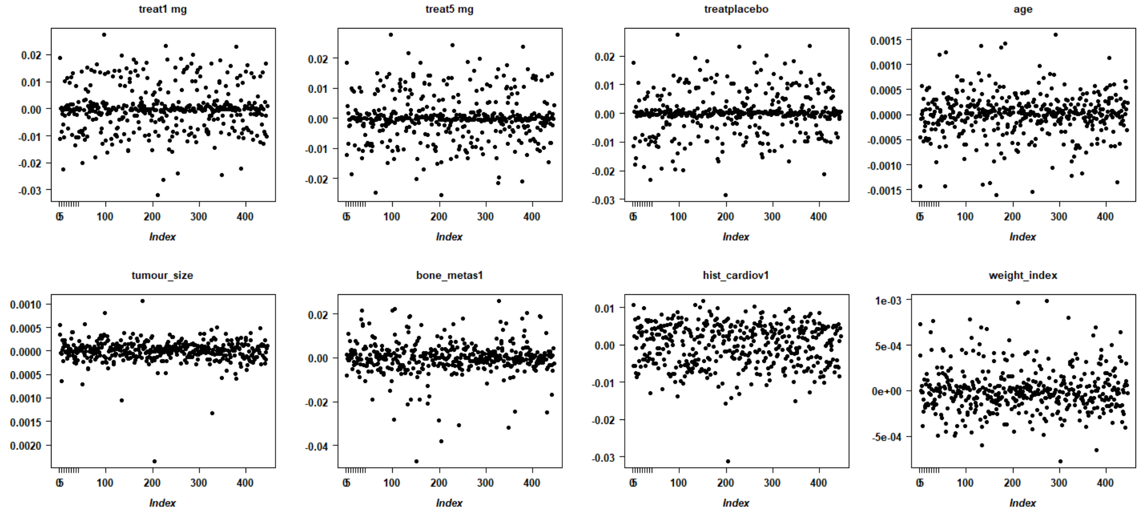


Figure 6: Dfbeta residuals.

6 Conclusions

After having carried out the different sections of the work, we can draw conclusions about the efficacy of the different treatments with a minimum of rigor. From the beginning of the work, when we estimated and plotted the survival functions, everything seemed to point to the 1 mg estrogen treatment being the most effective. However, at the beginning, with the non-parametric tools, it was difficult to determine which treatments followed the 1 mg treatment as the most effective.

Notwithstanding, thanks to the parametric model fitted in the fourth section of the work and the proportional hazard ratio model fitted in the fifth section, it has become clear that the most effective treatment is the 1 mg estrogen treatment, followed by the 5 mg estrogen treatment, then the 0.2 mg treatment and finally the placebo. On the other hand, we have been able to see that the difference between the last two treatments mentioned is not very great.

It should be noted that during the study we were surprised by the important effect of whether the patients had suffered from cardiovascular disease or not. We have been able to see that this is a variable clearly affects the survival of patients who receive this type of treatment for prostate cancer. Other variables, such as the presence of bone metastasis, are also crucial.

A Appendix

In this appendix we will add some figures and tables that we have not added in the work due to lack of space or because we consider them less important than those that have been added.

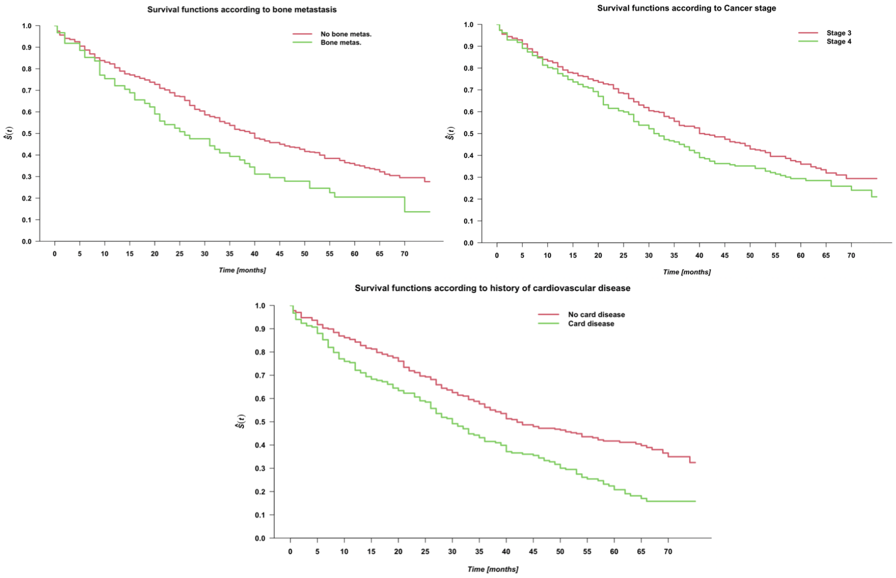


Figure 7: Survival functions according to (top left, top right and bottom) the bone metastasis, the cancer stage and history of cardiovascular diseases.

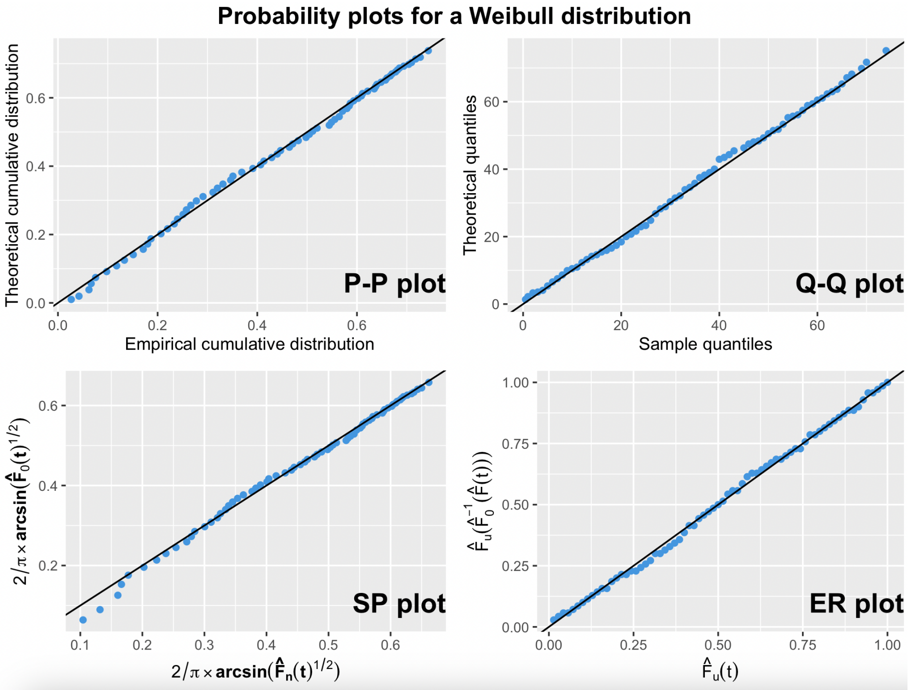


Figure 8: Probability plots for a Weibull distribution.

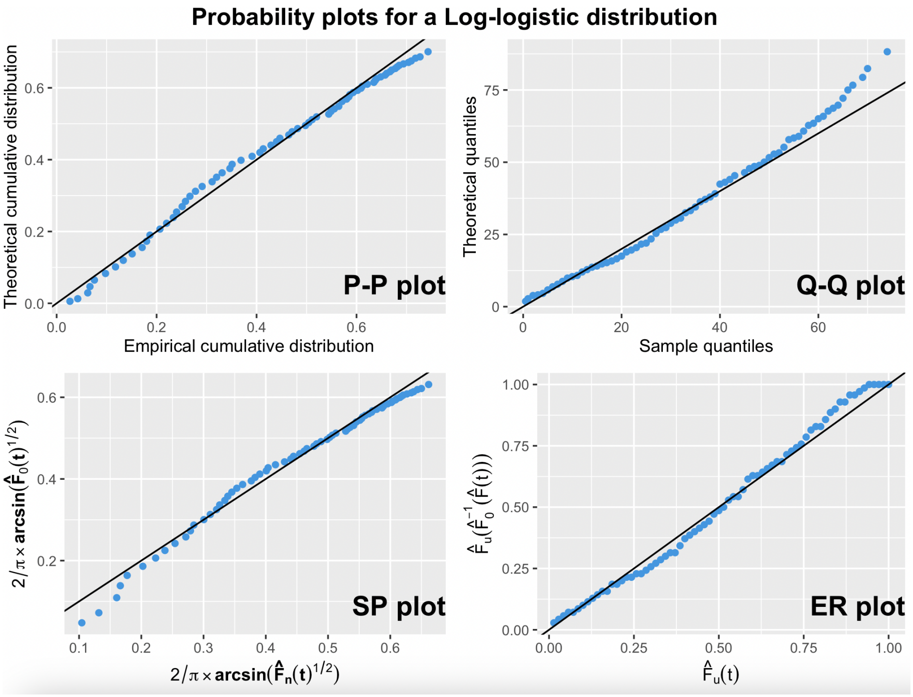


Figure 9: Probability plots for a Log-logistic distribution.

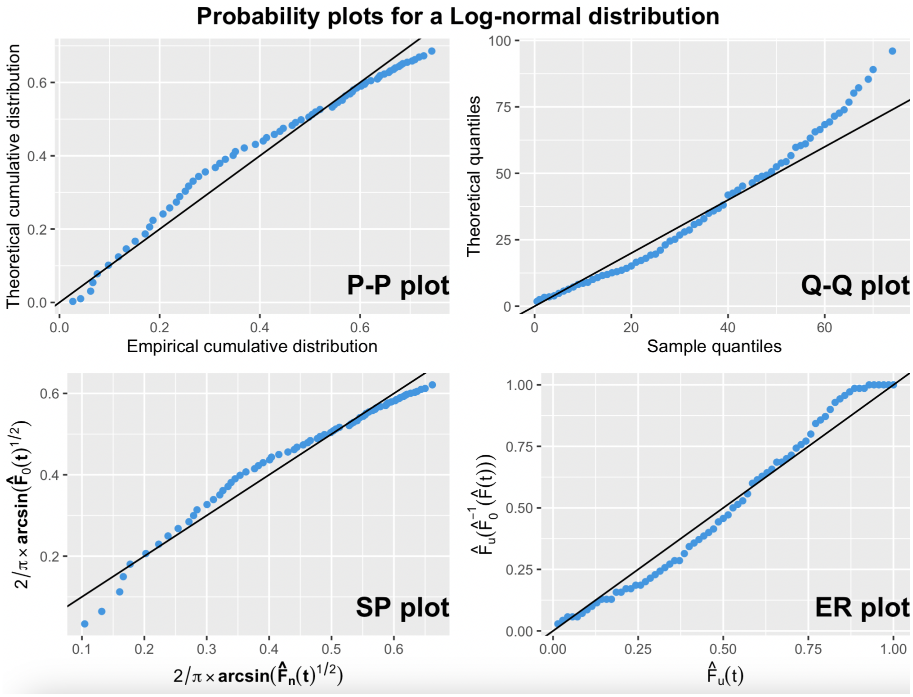


Figure 10: Probability plots for a Log-normal distribution.