Prediction of survival time of prostate cancer patients using Cox regression

Martina Kaponen
Prediction of survival time of prostate cancer patients using Cox regression

Martina Kaponen

June 2018
Abstract

Prostate cancer is the most common form of cancer in Sweden and the most common death related cancer form among men in Sweden. Because of that it is interesting to study the disease course and how different treatments effect the survival time. This thesis investigates the last part of the disease course to provide a basis for developing and use of new treatments. Using information about patients with prostate cancer from several national health care registers, Cox regression models were created for estimating the survival times. Biggest impact on survival time had treatment, age and whether metastases occurred when the patient were considered castration resistant.

Medical terms
PC Prostate Cancer
AA Antiandrogens
GnRH Gonadotropin-Releasing Hormone
PSA Prostate Specific Antigen
RP Radical Prostatectomy
RT Radio Therapy
CR Castration Resistant
DT Deferred Treatment
Orch Orchiectomy

1 Introduction

The most common form of cancer in Sweden, with over 10 000 new diagnoses at 2015, is prostate cancer, where it most commonly affects elderly men[2]. Common ways of treating prostate cancer are with surgery (RP), radio therapy (RT), orchiectomi, hormones (GnRH or/and AA), or a combination of them. RP and RT is a form of curative treatment in hope to be totally cured from the disease. One way to determine in which phase of the disease course a patient is and how the patient is responding to the treatment, is with a blood sample. The blood sample shows the amount of the prostate specific antigen (PSA), which is supposed to be low or at least non increasing if the patient responds well [2]. If the PSA values are rising after a curative treatment one way to slow down the disease course is to use GnRH. At some point that usually loses its effect and the PSA values start rising again. This is considered as an incurable state of the prostate cancer and the patient is said to be castration resistant. The normal definition of castration resistant is that, when the patient still is on medication, two consecutive samples of PSA are rising and are above 2µg/l [2]. Today there exist some new medication to use when a patient is CR, therefore more studies on this point of the disease course is needed. The purpose of this thesis is to investigate if there is any difference between the time from CR to death depending on initial treatment, age, metastases, stage group at diagnosis and region.
2 Methods and materials

2.1 Data

To be able to follow a patient over the entire treatment for prostate cancer the dataset that is used in this study is based on different national health care registers, like the national patient register, prescribed drug register, the cause of death and the national prostate cancer register. The total number of patients that were diagnosed with prostate cancer in these datasets were 12330 st. The datasets were merged together to create a new dataset. Only the PSA-samples that were taken after date of diagnosis were kept. To keep track on which PSA-samples that were affected by which treatment columns with only ones and zeros were created, one for each kind of treatment except for GnRH, which were given two columns. One for dependence after treatment including described daily dose plus thirty days and another one for dependence after treatment and described daily dose plus ninety days. The described daily dose is the number of days the treatment is supposed to be used and have an effect. Further, patients that only lived in Dalarna, Gävleborg, Uppsala, Värmland or Örebro during the time period of 2006-2015 were included in the study, since it is only then the whole treatment history is possible to follow. The timelap where there exists complete data differs between the regions and because of that the time lap is from 2006-01-01 to 2012-12-31 for patients that lived in Dalarna. 2006-01-01 to 2013-12-31 for Örebro and Värmland since the prescribed drug register was not complete. For the last two regions, Uppsala and Gävleborg the time lap is from 2006-01-01 to 2014-12-31. Everyone was diagnosed with PC and only PSA samples after date of diagnosis is considered.

There were 374 patients that were excluded even though they fulfilled all the criteria, but since the first expedition date of GnRH for these patients was between 2005-07-01 and 2006-01-01 it is not possible to be completely sure of when the first date of GnRH actually appeared. This is because the prescribed drug register only captures data from 2005-07-01. Since the patients that still are in the study do not have any prescription of GnRH between 2005-07-01 and 2006-01-01, the assumption is that their first expedition date is after 2006-01-01.
2.2 Castration resistant prostate cancer

The definition of castration resistant is defined by [2] as the following:
If the plasma testosterone value is less than 1.7 nmol/l and at least one of the following occurs the
patient is said to be castration resistant.

- Local progress valued by palpation or imaging.
- Newly arrived or growing metastases
- The PSA value is over $2\mu g/l$ and for two consecutive PSA samples, with at least one week
  apart, has rised with $1\mu g/l$.

Since the dataset only contained information about the PSA values, that was the property to
consider. Date of castration means the date that has the third property from above. To be a
part of the study the patients had to have had at least five PSA samples after 2006-01-01, and if
treated with radio therapy or radical prostatectomy, the considered castration date had to be after
this treatment. Additionally, the date of castration needed to either be affected by orchiectomy or
GnRH. If treated with orchiectomy, the day of the treatment only needed to occur before the date
of castration, but if treated with GnRH, the time between the most recent expedition date prior to,
and the castration date itself was allowed to be at most the defined daily dose plus 30 days. Also
the date of PSA-nadir was found in a similar way, which is the lowest PSA-value that occur before
the castration date and is affected by the same treatments as the date of castration.

![Figure 2: PSA-samples from patients treated with 2.1 and 2.2](image-url)
<table>
<thead>
<tr>
<th>Event</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orchietomy</td>
<td>grey</td>
</tr>
<tr>
<td>AA</td>
<td>purple</td>
</tr>
<tr>
<td>GnRH</td>
<td>red</td>
</tr>
<tr>
<td>PSA-nadir</td>
<td>black</td>
</tr>
<tr>
<td>CR</td>
<td>green</td>
</tr>
</tbody>
</table>

Figure 2 shows plots of PSA-values from two different patients, that were treated in different ways. The vertical lines represent events during the treatment course. To consider is that the time used for GnRH and AA is the expedition date, which means that it is impossible to be sure of the exact time of the treatment. Probably it is the same day or just a couple of days after this date. Date of PSA-nadir and CR comes from the methods described above. In figure 2.1 it is clear that after the orchietomy the PSA-values dropped for almost a year until it reached PSA-nadir and then, after the patient is considered castration resistant there is one expedition date of AA, this could be due to the definition of castration resistant that depends on more than just the PSA-values, but here not taken into account. In figure 2.2 the patient has two peaks before considered castration resistant. The first one is treated with AA and the second time the PSA-values increased GnRH was prescribed. PSA-nadir and castration resistant date is shown as before. In this case GnRH is continued to be prescribed even though the patient is considered as castration resistant, using the assumption of CR as described above.

### 2.3 Methods

The survival function is a monotone non increasing function that describes the time to an event, which starts in one and decreases to zero. This event can be, as the name suggest, death but also a lot of other things like appearance of a new tumor or recovery from a disease. The survival function is not only used in medical research, many other fields use this kind of function to describe the time to an event to occur, like time to failure of a component in a system. In theory the survival function is a smooth curve but in practice it is more of a step function depending on the time for observation, which could be for example hours, days, weeks or months.

If $X$ is a continuous random variable the probability of surviving, or the event to not occur, beyond time $x$, is defined by

**Definition. Survival function**[1][3]

$$S(x) = P(X > x)$$

where

$$S(x) = 1 - F(x) = 1 - P(X \leq x) = \int_x^\infty f(t) dt$$

where $f(t)$ is the density function and $F(x)$ is the cumulative distribution function of the continuous random variable $X$.

An important part of survival analysis is censored data, this is when some of the information is unknown. It could be that the event did not occur during the observation time, the participant dropped out of the study or was included at a later stage. But to take advantage of as much information as possible it is better to include these data even though they are not complete, than to exclude them, as described in [4]. The example below is a simplified version of an example in [5].
Example 1. Assume that the survival times observed in a study for simplicity are, 1, 2*, 3*, 6, 7, 8, 11, 15 and 19 days. The star marked times are censored, they dropped out of the study before the event occurred. Then the naive interpretation to estimate the probability of surviving beyond 7 days will be to exclude the censored and then get the following

\[ S(7) = P(X > 7) = \left(1 - \frac{1}{7}\right) \left(1 - \frac{1}{6}\right) \left(1 - \frac{1}{5}\right) = \left(\frac{6}{7}\right) \left(\frac{5}{6}\right) \left(\frac{4}{5}\right) = \frac{4}{7} \sim 0.57 \]

To get a more accurate probability measure and include as much information as possible it is preferable to include the censored data. This is done by conditioned probability, the probability to survive during time step \( x_i \) provided that they survived until time step \( x_{i-1} \). As shown below the censored data is included for the time it is relevant and then excluded.

Example 1, cont. Including the censored data gives the estimated probability of surviving more than 7 days

\[ S(7) = P(X > 7) = \left(1 - \frac{1}{9}\right) \left(1 - \frac{1}{6}\right) \left(1 - \frac{1}{5}\right) = \left(\frac{8}{9}\right) \left(\frac{5}{6}\right) \left(\frac{4}{5}\right) \sim 0.59 \]

Hence the probability of surviving increased when including the censored data.

To get more information about the survival function it is often interpreted with the hazard function or the hazard ratio. The hazard ratio is a measure of how often an event occurs in one group compared to another group. If the ratio is one, there is no difference between the groups. The groups could be for example different treatments or one with treatment and the other one with placebo. If \( X \) is a continuous random variable, the hazard ratio is the probability that at a specific time \( x \) the event has not occurred yet. In other words, if the event has not occurred then the hazard rate is the probability that the event will occur in the next time step divided by the time step. The hazard ratio is non-negative for all \( x \) and the limit of the integral \([0, \infty]\) is infinity.

**Definition.** Hazard function

\[ h(x) = \lim_{\Delta x \to 0} \frac{P(x \leq X < x + \Delta x | X \geq x)}{\Delta x} \]

\[ H(x) = \int_0^x h(u) du \]

The Hazard function and the survival function has the following relations.

\[ h(x) = \frac{f(x)}{S(x)} \]

\[ S(x) = \exp^{-H(x)} \]
2.4 Cox regression model

Cox regression is one of the most common ways to compare survival times depending on predictors, $(x_1 \ldots x_n)$. The predictors could be either continuous or categoric.

$h_0$ is the baseline hazard and refers to the outcome if all the predictors are set to 0, i.e. the model does not have any predictors.

**Definition.** Cox proportional hazard model. [1]

\[ h(t) = h_0(t) \cdot \exp(b_1 x_1 \ldots b_n x_n) \]

\[ \ln \left( \frac{h(t)}{h_0(t)} \right) = b_1 x_1 + \ldots + b_n x_n \]

This model is also called Cox proportional hazard model since the ratio of the hazard rates $h(t)/h_0(t)$ is independent of $t$. If the covariates are dependent on time this model does not hold. Then one solution is to introduce another covariate as a function of time, the problem is to find that function [7].

2.5 Likelihood ratio test

The likelihood ratio test is a way of comparing two hypotheses. As the name suggest the likelihood ratio between the two hypotheses is compared to some reference distribution by taking the ratio of the likelihood functions and comparing that to some reference distribution. If the likelihood ratio is above the critical value the null hypothesis will be rejected [1][6].

\[ \chi^2_\alpha(r) < -2 \ln \frac{L_0}{L_A} \] reject the null hypothesis

\[ \chi^2_\alpha(r) > -2 \ln \frac{L_0}{L_A} \] do not reject the null hypothesis

Where $\alpha$ is the desired quantile of the $\chi^2$ distribution, $r$ = degrees of freedom = predictors of the alternative model - predictors of the null model and $L_0$ and $L_A$ are the likelihoods of the respective model.

3 Results

The total set to work with contained 8808 different patients and 119516 observations, where observations is the sample times of PSA-values. The data set contains the patients which were diagnosed with prostate cancer and lived in the right region. Since most patients have been treated with a combination of two or more treatments, they could occur in multiple rows in table 1. The column Treatment refers to the current treatment and the column History refers to treatments that occurred before the current treatment. #Observation is the total number of PSA-samples taken of patients with the respective history and treatment. #Patients is the number of patients that at some point had this combination of treatment and history and Samples/year is the mean amount of PSA-samples taken each year. The time lap for calculating this mean is from the date of one treatment until another treatment occurred, time of death or 2014-01-01.
Table 1: Distribution of observations, patients and samples depending on treatment

<table>
<thead>
<tr>
<th>History</th>
<th>Treatment</th>
<th>#Observations</th>
<th>#Patients</th>
<th>Samples/year</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td>33686</td>
<td>6579</td>
<td>2.2718</td>
<td></td>
</tr>
<tr>
<td>RP</td>
<td>2906</td>
<td>336</td>
<td>2.6938</td>
<td></td>
</tr>
<tr>
<td>RT</td>
<td>4198</td>
<td>825</td>
<td>3.6852</td>
<td></td>
</tr>
<tr>
<td>AA</td>
<td>3651</td>
<td>1229</td>
<td>3.6883</td>
<td></td>
</tr>
<tr>
<td>RP/ RP+RT</td>
<td>GNRH/ORCH</td>
<td>2434</td>
<td>146</td>
<td>4.5421</td>
</tr>
<tr>
<td>RT</td>
<td>GNRH/ORCH</td>
<td>2909</td>
<td>749</td>
<td>2.5565</td>
</tr>
<tr>
<td>DT/ AA</td>
<td>GNRH/ORCH</td>
<td>40910</td>
<td>3640</td>
<td>2.8945</td>
</tr>
<tr>
<td>Total</td>
<td>GNRH/ORCH</td>
<td>78649</td>
<td>6478</td>
<td>2.6785</td>
</tr>
</tbody>
</table>

All individuals in table 2 are considered castration resistant due to the definition above. Group one is referring to all patients that are treated with GnRH or orchiectomy, are CR and with at least five PSA samples, group two is a subset of group one with the additional condition that first date of expedition for GnRH or date of orchiectomy is after 2006-01-01.

Table 2: Distribution of patients in respective predictor group

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at CRPC</td>
<td>1. ≤ 65</td>
<td>193(15.1%)</td>
</tr>
<tr>
<td></td>
<td>2. 66-75</td>
<td>484(37.9%)</td>
</tr>
<tr>
<td></td>
<td>3. 76-85</td>
<td>498(39%)</td>
</tr>
<tr>
<td></td>
<td>4. ≥ 86</td>
<td>102(8%)</td>
</tr>
<tr>
<td>Treatment history</td>
<td>DT</td>
<td>849(66.5%)</td>
</tr>
<tr>
<td></td>
<td>Cur</td>
<td>71(5.6%)</td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td>36(2.8%)</td>
</tr>
<tr>
<td></td>
<td>GnRH</td>
<td>321(25.1%)</td>
</tr>
<tr>
<td>STG Group at diagnosis</td>
<td>1. Low risk</td>
<td>49(3.8%)</td>
</tr>
<tr>
<td></td>
<td>2. Intermediate risk</td>
<td>114(8.9%)</td>
</tr>
<tr>
<td></td>
<td>3. High risk</td>
<td>331(25.9%)</td>
</tr>
<tr>
<td></td>
<td>4. Regionally metastatic</td>
<td>194(15.2%)</td>
</tr>
<tr>
<td></td>
<td>5. Distant metastases</td>
<td>578(45.3%)</td>
</tr>
<tr>
<td></td>
<td>6. Missing Data</td>
<td>11(0.9%)</td>
</tr>
<tr>
<td>Region</td>
<td>Värmland(S)</td>
<td>265(20.8%)</td>
</tr>
<tr>
<td></td>
<td>Örebro(T)</td>
<td>184(14.4%)</td>
</tr>
<tr>
<td></td>
<td>Uppsala(U)</td>
<td>358(28%)</td>
</tr>
<tr>
<td></td>
<td>Dalarna(W)</td>
<td>194(15.2%)</td>
</tr>
<tr>
<td></td>
<td>Gävleborg(X)</td>
<td>276(21.6%)</td>
</tr>
<tr>
<td>Metastases</td>
<td>M1 at CR</td>
<td>375(29.4%)</td>
</tr>
<tr>
<td></td>
<td>No M1 at CR</td>
<td>902(70.6%)</td>
</tr>
</tbody>
</table>
Figure 3: Survival curves of overall survival for group 2
Figure 4: Survival curves of cure specific survival for group 2

(a) Age at CRPC
(b) Metastases at date of CRPC
(c) Treatment history
(d) Region
Figure 5: Survival curves for patients in treatment group DT depending on stage group at diagnosis

(a) Cure specific survival

(b) Overall survival

Figure 6: Proportion of cure specific and overall survival times
Figure 3 and 4 show survival curves depending on age, metastases, stage group at diagnosis and initial treatment. There are very small differences between figure 3 and 4 which can be explained by figure 6, which shows the difference between the amount of overall survival and cure specific survival. There is a total of 626 cure specific and 73 of other causes and therefore 699 in the overall survival group.

### 3.1 Cox regression model

coxph() is an inbuilt function in R to create a cox regression model of the form

\[ h(t) = h_0(t) \cdot \exp(b_1x_1 \ldots b_nx_n) \]

where \( b \) is the coefficient of each predictor \( x \). The model created for group 2 with this function has coefficients for age, treatment, metastases and region. A negative coefficient has negative impact on the the hazard and hence a positive impact on the survival time. The result in table 3 can be compared to figure 4 and 3. Table 4 shows the coefficients from a subset of group 2, where all patients belong to the treatment group DT. This group was under what is called watchful waiting during the first period of time after they were diagnosed with prostate cancer. This in general means that they were in an early stage of the disease course and no treatment was needed in this stage. Since the stage group at diagnosis can be assumed to have a high correspondence to which treatment the patient receives, this division of models was made. The complete outputs for these models are presented in appendix A.1.

If only consider table 3, group 2, the treatment group AA is used as the baseline hazard and all the other treatments have a negative coefficient, this tells us that patients with another treatment than AA have a negative input on the hazard, which means that they have a positive input on the total survival time. The patients that received curative treatment have the lowest coefficient and hence the longest survival times.

In tables 3 and 4 the coefficient for age is not divided in age groups as in the figures 3 and 4. In both cases however the age has a negative impact on the survival time. What can be seen is that the overall survival is more dependent on age than the cure specific survival. This is not a surprising result, since at a higher age other causes of death are more common. Both tables also have similar results for the impact of metastases and region. Where occurrence of metastases before CR will have a negative impact of the survival time. Regarding region there are no big differences, but there is a slightly higher survival time for patients that lived in Uppsala compared to the other regions.

The results above can be compared to the output of the survival function, survfit() in R, for the cure specific survival times, which were plotted in figure 4 and can be found in appendix A.2. The median survival times presented there consider number of days from CR to death. For the age groups the median for the patients over 86 were 581 days comparing to the age group from 66 to 75 year where it was 710. To consider here though is that the 95 percent confidence interval for all of the age groups are quite similar. Depending on if metastases existed before the date of CR a bigger difference in survival time is shown, if metastases existed before CR the survival time in days were 448 comparing to 784 days for the ones without metastases at that point. Here the confidence intervals were totally separated which could indicate that this is an important factor to take into account. As for the age groups, the stage groups and regions did have some small differences in median but the confidence intervals were similar and no big conclusions should be considered. The treatment history on the other hand have similar results for both deferred treatment and patients treated initially with GnRH, around 650 days. But patients treated with AA had only 383 days compared to the ones treated with some of the curative methods who had 1057 days as a median.

### 3.2 Likelihood ratio test of models

A likelihood ratio test can help to determine whether the model is a good fit or if another approach would generate a better model. The output of coxph() includes the likelihood ratio test where the
null model is the assumption of no predictors at all and the alternative model is the one created from `coxph()`. In other words the likelihood ratio test value presented in the output compares if the model created is better than a model without any predictors. For the model of the cure specific data in table 3 we have the value of the likelihood ratio test to 86.08 on 9 degrees of freedom, where the degrees of freedom comes from the number of predictors. This value is compared to $\chi^2_{0.95}(9) = 16.92 < 86.08$, hence we can reject the null model and conclude that the alternative model is a better fit. This is true for all the models presented in table 3 and 4.

Another approach is to model this dataset without the predictors about region, the output for that approach is found in appendix A1. The likelihood ratio test in that case is 72.51 with 5 degrees of freedom, since there are four less predictors. In this case $\chi^2_{0.95}(5) = 11.07 < 72.51$, hence this model is also a good fit compared to the null model. But is the bigger model a significant better fit than this model? To determine that we can calculate what the value of the likelihood ratio test between these test would be.

The likelihood of the null model, without predictors, is represented with $L_0$. The likelihood of the smaller model, without region as a predictor is represented by $L_{A1}$ and the likelihood of the bigger model is represented by $L_{A2}$.

$$\lambda_1 = -2 \ln \frac{L_0}{L_{A1}} = -2 \ln L_0 + 2 \ln L_{A1} = 72.51$$
$$\lambda_2 = -2 \ln \frac{L_0}{L_{A2}} = -2 \ln L_0 + 2 \ln L_{A1} = 86.08$$

The assumption that $A_1$ is the new null model to compare with the alternative model $A_2$ gives that

$$\lambda = -2 \ln \frac{L_{A1}}{L_{A2}} = 2 \ln L_{A2} - 2 \ln L_{A1} =$$
$$= 2 \ln L_{A2} - 2 \ln L_{A1} + (2 \ln L_0 - 2 \ln L_0) =$$
$$= -2 \ln L_0 + 2 \ln L_{A2} - (2 \ln L_0 - 2 \ln L_{A1}) =$$
$$= \lambda_2 - \lambda_1 = 86.08 - 72.51 = 13.57 > \chi^2_{0.95}(9 - 5) = 9.488$$

Hence the null model (A1) can be rejected, the model with all predictors (A2) is a better fit.

As described above, one of the reasons to create a model of patients from the treatment group DT were to conclude if the stage group at diagnoses had any impact on the survival time. The likelihood ratio for the cure specific in table 4 is $\lambda_3 = 56.03$ with 6 degrees of freedom, and for a model which included the stage group at diagnosis the likelihood ratio is $\lambda_4 = 58.87$ with 9 degrees of freedom. To conclude if this bigger model is a significantly better fit than the one in table 4, similar computations as above results in:

$$\lambda_4 - \lambda_3 = 58.87 - 56.03 = 2.84 < \chi^2_{0.95}(9 - 6) = 7.815$$

Hence the null model should not be rejected and the smaller model without a predictor depending on which stage group the patient were in for the time of diagnosis should be considered.
<table>
<thead>
<tr>
<th>Overall hazard</th>
<th>Cure specific hazard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>coef</td>
</tr>
<tr>
<td></td>
<td>0.0222</td>
</tr>
<tr>
<td>Treatment</td>
<td>AA</td>
</tr>
<tr>
<td>Curative</td>
<td>-0.9227</td>
</tr>
<tr>
<td>GnRH</td>
<td>-0.6898</td>
</tr>
<tr>
<td>DT</td>
<td>-0.6187</td>
</tr>
<tr>
<td>Metastases</td>
<td>Not before CR</td>
</tr>
<tr>
<td>Before CR</td>
<td>0.7405</td>
</tr>
<tr>
<td>Region</td>
<td>Värmland</td>
</tr>
<tr>
<td></td>
<td>Örebro</td>
</tr>
<tr>
<td></td>
<td>Uppsala</td>
</tr>
<tr>
<td></td>
<td>Dalarna</td>
</tr>
<tr>
<td></td>
<td>Gävleborg</td>
</tr>
</tbody>
</table>

Table 3: Output from coxph() in R for group 2

<table>
<thead>
<tr>
<th>Overall hazard</th>
<th>Cure specific hazard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>coef</td>
</tr>
<tr>
<td></td>
<td>0.0170</td>
</tr>
<tr>
<td>Metastases</td>
<td>Not before CR</td>
</tr>
<tr>
<td>Before CR</td>
<td>0.8366</td>
</tr>
<tr>
<td>Region</td>
<td>Värmland</td>
</tr>
<tr>
<td></td>
<td>Örebro</td>
</tr>
<tr>
<td></td>
<td>Uppsala</td>
</tr>
<tr>
<td></td>
<td>Dalarna</td>
</tr>
<tr>
<td></td>
<td>Gävleborg</td>
</tr>
</tbody>
</table>

Table 4: Output from coxph() in R for the subset DT of group 2
4 Discussion

One of the biggest issues during this study were to determine if a patient should be considered as castration resistant or not. The only way to determine this from the dataset were by the PSA-values. For future studies it could be of interest to, in the same way as collecting the date of diagnosis, collect the date for when the patient is considered as castration resistant. That could give a better understanding on why a treatment ended or begun at a certain point. Considering the output from the survival function and the median survival times presented there, some conclusions about metastases and treatment could be done, like that the survival time probably is longer if metastases do not occur in an early stage of the disease course. In some way it could be tempting to draw the same conclusions depending on which treatment a patient was given. That if treated with AA the survival time is shorter than with a curative method for example. Even if the tables are showing that result, it does not say anything about why a certain treatment was chosen. The patients treated with AA could be older or have an more aggressive form of cancer than those treated with a curative method for example. To take into account what impact the stage group a patient belonged to at the time of diagnosis some comparisons were made with only patients from the group deferred treatment. This however did not show any significant results either.

References


### Appendices

#### A Output from R

##### A.1 Cox regression, coxph()

| Coef | Exp(coef) | SE(coef) | z | Pr(>|z|) |
|------|-----------|----------|---|---------|
| Age | 0.01562 | 1.01575 | 0.00610 | 2.561 | 0.010461 * |
| HbBeforeCR | 0.77735 | 2.17799 | 0.10957 | 7.109 | 0.000012 *** |
| region | -0.1106 | 0.9006 | 0.1480 | -0.635 | 0.525576 |
| regionU | -0.4822 | 0.6179 | 0.1488 | -3.340 | 0.000838 *** |
| regionX | -0.1504 | 0.8215 | 0.1509 | -1.045 | 0.292039 |

---

### 7.1 Cure specific survival

Figure 7: Deferred treatment, without the predictor stage group

### 7.2 Overall survival
8.1 Cure specific survival

Figure 8: Deferred treatment, including the predictor stage group

8.2 Overall survival

10.1 Cure specific survival

Figure 10: All treatments, without the predictor region

10.2 Overall survival
9.1 Cure specific survival

A.2 Survival funcion

The output from the cure specific survival functions plotted in figure 4.

Call: survfit(formula = Surv(time, censor == 1) ~ AgeGrp + treatment + MibefronR + region, data = Cisafter06)

n events median 0.95LCL 0.95UCL

CRPC$AgeGrp=66to75 335 232 710 610 800
CRPC$AgeGrp=76to85 324 234 648 569 762
CRPC$AgeGrp=geq86 71 50 581 487 846
CRPC$AgeGrp=leq65 144 110 672 579 859

Call: survfit(formula = Surv(time, censor == 0) ~ AgeGrp + treatment + MibefronR + region, data = Cisafter06)

n events median 0.95LCL 0.95UCL

CRPC$M1beforeCR=0 568 417 784 738 842
CRPC$M1beforeCR=1 306 209 448 385 502

Call: survfit(formula = Surv(time, censor == 0) ~ CRPC$STG_Group)

n events median 0.95LCL 0.95UCL

17
CRPC$STG_Group=1-2 109 75 696 587 866
CRPC$STG_Group=3 245 169 816 646 878
CRPC$STG_Group=4-5 509 375 625 555 712
CRPC$STG_Group=6 11 7 483 359 NA

Call: survfit(formula = SurvTime2to0 ~ CRPC$treatment)

<table>
<thead>
<tr>
<th>CRPC$treatment</th>
<th>n</th>
<th>events</th>
<th>median</th>
<th>0.95LCL</th>
<th>0.95UCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>36</td>
<td>20</td>
<td>383</td>
<td>288</td>
<td>502</td>
</tr>
<tr>
<td>Cur</td>
<td>58</td>
<td>31</td>
<td>1057</td>
<td>800</td>
<td>1485</td>
</tr>
<tr>
<td>Direct</td>
<td>162</td>
<td>133</td>
<td>609</td>
<td>543</td>
<td>805</td>
</tr>
<tr>
<td>DT</td>
<td>618</td>
<td>442</td>
<td>665</td>
<td>606</td>
<td>752</td>
</tr>
</tbody>
</table>

Call: survfit(formula = SurvTime2to0 ~ CRPC$region)

<table>
<thead>
<tr>
<th>CRPC$region</th>
<th>n</th>
<th>events</th>
<th>median</th>
<th>0.95LCL</th>
<th>0.95UCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>166</td>
<td>120</td>
<td>694</td>
<td>602</td>
<td>805</td>
</tr>
<tr>
<td>T</td>
<td>131</td>
<td>100</td>
<td>543</td>
<td>468</td>
<td>731</td>
</tr>
<tr>
<td>U</td>
<td>238</td>
<td>158</td>
<td>778</td>
<td>625</td>
<td>866</td>
</tr>
<tr>
<td>W</td>
<td>137</td>
<td>115</td>
<td>603</td>
<td>549</td>
<td>771</td>
</tr>
<tr>
<td>X</td>
<td>202</td>
<td>133</td>
<td>669</td>
<td>550</td>
<td>805</td>
</tr>
</tbody>
</table>

Output of the cure specific survival function for stage groups in the treatment group deferred treatment.

Call: survfit(formula = SurvTime2to0 ~ DT$STG_Group)

<table>
<thead>
<tr>
<th>CRPC$STG_Group</th>
<th>n</th>
<th>events</th>
<th>median</th>
<th>0.95LCL</th>
<th>0.95UCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>80</td>
<td>57</td>
<td>607</td>
<td>552</td>
<td>796</td>
</tr>
<tr>
<td>3</td>
<td>190</td>
<td>135</td>
<td>816</td>
<td>640</td>
<td>913</td>
</tr>
<tr>
<td>4-5</td>
<td>341</td>
<td>246</td>
<td>633</td>
<td>546</td>
<td>738</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>4</td>
<td>483</td>
<td>150</td>
<td>NA</td>
</tr>
</tbody>
</table>

B Code R

B.1 Castration resistant

kastdate= c(0); kastlopnr=c(0); nadirdate=c(0); PSA nadir=c(0)
i=1; j=1; m=1; n=1; p =1
nolopnr=count(WorkData$lopnr)[,2]
while(i<length(WorkData$lopnr)){
  noloops=nolopnr[i]
  firstDate= "2006-01-01"
  if(WorkData$region[i]== "W"){
    lastDate = "2012-12-31"
  }else if((WorkData$region[i]== "T") | (WorkData$region[i]== "S")){
}
lastDate = "2013-12-31"
} else {
    lastDate = "2014-12-31"
}
for (k in (i : (i+noloops - 3))) {
    if (WorkData$Provdatum[k] > firstDate &
        WorkData$Provdatum[k+2] < lastDate) {
        m=m+1
        if (((!is.na(WorkData$RPdate[k])) &
            WorkData$RPdep[k] == 1) | is.na(WorkData$RPdate[k])) &
            (((!is.na(WorkData$RTdate[k])) & WorkData$RTdep[k] == 1)) |
            is.na(WorkData$ORCHdep[k]) &
            (WorkData$ORCHdep[k] == 1) | is.na(WorkData$ORCHdtm[k]) &
            (WorkData$ORCHdep[k] == 1 | WorkData$gnrhdep1[k] == 1)) {
            n=n+1
            if (((WorkData$PSA[k] < WorkData$PSA[k+1]) &
                (WorkData$PSA[k+1] < WorkData$PSA[k+2]) &
                ((WorkData$PSA[k+2] - WorkData$PSA[k]) > 1) &
                WorkData$PSA[k+2] > 2)) {
                kastlopnr[p] = WorkData$lopnr[k]
                kastdate[p] = WorkData$Provdatum[k+2]
                PSAnadir[p] = min(WorkData$PSA[i : (k+2)])
                index = tail(which(WorkData$PSA[i : (k+2)] == PSAnadir[p]), 1)
                nadirdate[p] = WorkData$Provdatum[(i+index - 1)]
                p=p+1
                break
            }
        }
    }
}
if (i>100 & i <150) {print(i)}
if (i>1000 & i<1050) {print(i)}
if (i>5000 & i <5050) {print(i)}
if (i>10000 & i <10050) {print(i)}
if (i>15000 & i <15050) {print(i)}
if (i>20000 & i <20050) {print(i)}
flush.console()
i=i+noloops
j=j+1
}