2012•2013
FACULTY OF SCIENCES
Master of Statistics: Biostatistics

Masterproef

Analysis of prostate cancer data at the presence of latent factor: Estimation of the treatment effect

Promotor :
Prof. dr. Ziv SHKEDY

Promotor :
Dr. PUSHPIKE THILAKARATHNE

Afroza Polin

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

universiteit hasselt
KNOWLEDGE IN ACTION

Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt
Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek

universiteit hasselt | Maastricht University

2012•2013
FACULTY OF SCIENCES
Master of Statistics: Biostatistics

Masterproef

Analysis of prostate cancer data at the presence of latent factor: Estimation of the treatment effect

Promotor :
Prof. dr. Ziv SHKEDY

Promotor :
Dr. PUSHPIKE THILAKARATHNE

Afroza Polin

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

universiteit hasselt | Maastricht University

Abstract

In medical and epidemiological research a common technique is to categorize a continuous variable before evaluating its prognostic impact on the clinical trial outcome of interest. This study is related to a clinical trial; which was designed to test the effectiveness of a new treatment compared to a currently used treatment for prostate cancer. In this study, we focused on finding a dichotomous latent factor based on the observed risk factors - pain score and PSA level, and then estimating the treatment effect. The patients having low pain and PSA level lower than the estimated threshold were considered as asymptomatic patient and the others are symptomatic patients. We used Cox(1972) model to estimate the effect of the covariates on survival time. We have applied different approach to estimate the threshold for PSA. In first approach, we estimated the best cutoff point as the PSA with maximum differences in treatment effect between asymptomatic and symptomatic patients. This estimated cutoff separates the patients with significantly different treatment effect. Separation of the patients in a better way does not necessarily indicate the best fit of the model to the data. In the second approach, we estimated the threshold by maximizing goodness of fit components; likelihood, c-index and concordance probability estimate. These estimates of cutoff failed to show significant differences in treatment effect between two groups of patients. The results of the parameter estimates in external validation were not as good as the train set,when threshold was estimated based on maximum differences of treatment effect. Whereas, the estimated threshold in the second approach, based on goodness of fit, had shown more stable estimates for new data set. Thirdly, We used another technique, Bayesian change point model for estimating the threshold and the regression parameters in Cox model. The threshold and the effect estimated in this method were closer to the estimate using goodness of fit components.

Finally it was concluded based on the second and third approach that, the effect of the treatment was not different in asymptomatic and symptomatic patients.

Keywords: Cutoff point, latent factor, Bayesian change point model.

Contents

List of Tables

Acknowledgements

First and foremost I would like to offer my sincerest gratitude to my supervisor, Professor Dr. Ziv Shkedy, who has supported me throughout my thesis with his patience and knowledge. At the same time I would like to express my gratitude to Dr. Pushpike Thilakarathne, Dr. Suzy Van Sanden and Dr. Joris Diels from Janssen to guide me during the thesis period.

I would like to express my gratitude to the Professors guided us in different courses, which provide me sufficient knowledge and confidence to do the thesis.

Also I would like to say thanks to my classmates, especially my friendly and cheerful group mates Victoriah, Mhiret and Hung for their cordial support during class and group meetings.

My sincere also goes to all the Hasselt University staffs. Especially, Mrs. Martine Machiel, she has supported us all with her sincerity and smiles.

My deepest gratitude always to my family members for their love and continuous support.

Finally, I thank the Almighty to do all these things successfully.

1 Introduction

Prostate cancer is a form of cancer that develops in the prostate, a gland in male reproductive system. Rates of detection of prostate cancers are higher in developed countries compare to developing countries. Globally, prostate cancer is the sixth leading cause of cancer related death in men [1].

Prostate-specific antigen (PSA) is a protein made by prostate tissue; can be detected in blood. Cancer cells make excessive amounts of this protein, so usually rising PSA levels are used as a signal of having something wrong with the prostate. PSA threshold for suspecting prostate cancer varies with age and race of the patients. Also the annual rate of change in PSA level is considered as an important factor[2]. Generally the higher PSA level and the faster rate of increasing the PSA, indicates the presence of more cancer cell in the body. But this is not always true. Sometimes PSA level may not be increased in the presence of prostate cancer.

Different types of treatments are available for prostate cancer patients. Some treatments are currently used; standard treatment and some are being tested in clinical trials. Prostate cancer could be very aggressive or very slow-growing. The Prognosis and treatment options depend on the stage of the cancer (PSA level, grade of tumor, whether the cancer has spread to other place in the body), patient's age and some other related factors[3].

Dichotomizing continuous covariates is a common practice in medical and epidemiological research for both clinical and statistical reasons. From a clinical point of view, binary covariates may be preferred for classifying high or low risk/response set. From statistical point of view, binary covariates may be preferred for simple interpretation of common effect measures from statistical model such as odds ratios and relative risk. The choice of a cut point to dichotomize a continuous variable needs attention. Most often biological knowledge about a risk factor or the

results in earlier studies is used. If well established cut points are not available, then statistical techniques are required to determine the cutoff point. Two broad categories of statistical methods to select the cutoff point are data-oriented and outcome-oriented. Data oriented method involves - using mean or any certain quantile as cutoff point. And the outcome oriented methods provide a value of a cut point that correspond to the most significant relation with outcome. Generally, the outcome-oriented methods are preferred compared to data-oriented methods. Different statistical methods for finding the candidate cutoff point has been applied in literature, such as *minimum p-value* approach or alternatively *the maximum statistic* approach . Kim (2004) [4] used profile likelihood to find cut point to dichotomize a continuous covariate in Cox model. Williams et. al. (2006) [5] used maximum differences between the two dichotomous group using the log-rank statistic for time-to event outcome. Some other criteria for choosing an optimal cutoff point have been applied, like maximum effect size, maximum precision of estimates.

In this study, the overall survival endpoint was compared between the patients receiving two treatment: a standard treatment available in market and a new treatment. The patients were not homogeneous in terms of risk factors. So the treatment effect should be estimated considering the effect of the other considered covariates; pain score and PSA level. The latent covariate group was defined as a function of these two observed covariates, patients having low pain and PSA lower than the threshold would be in asymptomatic group, otherwise in symptomatic group. Therefore, the ultimate goal of this study was to estimate the optimal cutoff point for PSA level and estimate the treatment effect.

In this report we have described the data set used for the analysis in section two. The optimum cutoff point for PSA was estimated using different outcome oriented objects - difference in hazard ratio between subgroup patients and goodness of fit components. And the effect of the covariates (treatment and group) was estimated using Cox model with the selected cutoff. These

methodologies have been described in section three. In section four application of these methods to the data set and selected results have been presented. In section five we have done a simulation study. In chapter six change point modelling technique has been applied in Bayesian frame work for the considered Cox model. And in the final chapter we discuss the results and make conclusion based on it.

2 Data

Data used in this report is from a clinical trial conducted on prostate cancer patients in different countries. 1088 patients were observed to assess the effectiveness of a new prostate cancer treatment, compared to a standard treatment. Two end points were considered in the main trials; overall survival and progression free survival. In this study we considered only the overall survival endpoint. About half of the patients (49.6%) received the experimental drug and the other half was treated with the standard drug. The covariates considered in the study were the treatment (1=new treatment, 0=standard treatment), pain score (1=low pain and 0=high pain) and PSA level in observed count (ng/mL).

For overall survival endpoint, 434 events were observed during the study period of 1088 days and 654 patients were censored. Table 1 shows, among 546 patients receiving the experimental treatment, 200 (36.63%) were experiencing the event death during the study period. While among 542 patients receiving the standard treatment, 234 (43.17%) were experiencing the event death in the study duration.In addition, 253 (35.33%) out of 716 patients, reported lower pain had the event (death), and 181 patients out of 372 (48.66%) having higher pain died during the study time. In an overall observation, censored patients had more wider range of PSA values (0.04, 6606.44) compared to PSA values (2.17, 3927.43) of the patients had died.

Table 1: Overall Survival Endpoint

	Experimental Treatment	Standard Treatment
Number of Patients	546	542
Number of events	200 (36.63 %)	246 (43.17 %)
Median Survival Time	1074 days	917 days

Kaplan-Meier estimate of the survival function for the two treatment groups are in figure 2, indi-

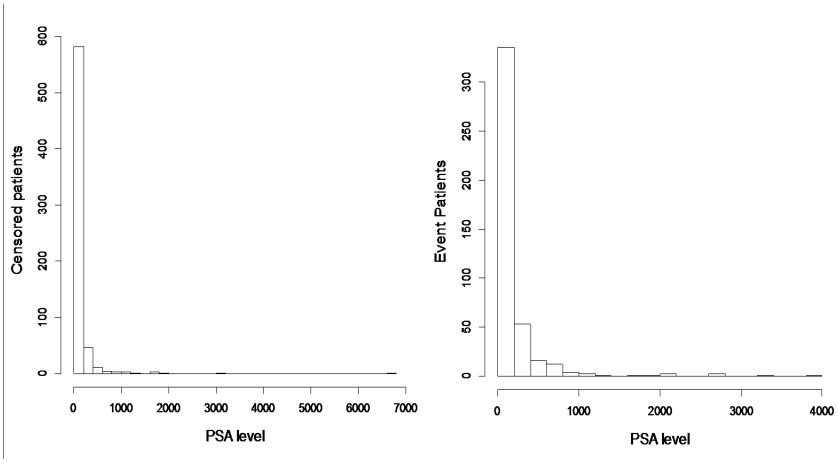

Figure 1: PSA level among censored and event patients

cates the higher chance of surviving for the patient using the experimental treatment as compare to the patients using the standard treatment.

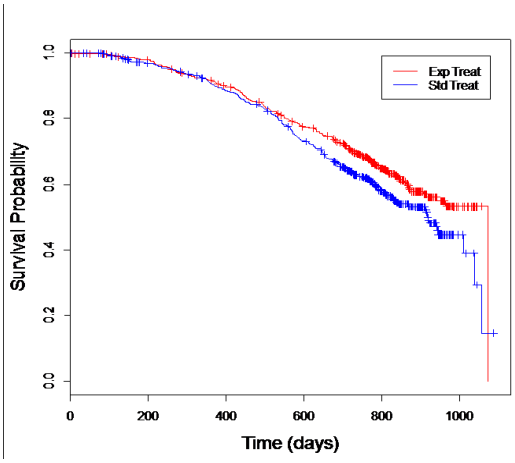

Figure 2: KM survival curves for two treatment groups

3 Methods

This study is related to a clinical trial which was designed to check the effectiveness of a new treatment compared to a standard one interms of overall survival and progression free survival. In this study analysis we used only the overall survival endpoint. The study objective was to create a latent group variable as a function of the observed covariates - pain score and PSA level, and estimate the treatment effect. Formation of group variable was mainly related to select cutoff point of PSA. In this section, we discuss the methods of estimating the best cutoff point and fit Cox model to estimate the treatment effect after adjusting the other covriates effect.

3.1 Basic Survival Model

The We used Cox (1972) model for analyzing the effect of covariates on survival. Treatment was the main interest covariate in this study. The Cox model for assessing only the treatment effect, ignoring the effect of other covariates

$$\lambda(t|\mathbf{Z}) = \lambda_0(t)\exp(\beta\text{treatment}) \tag{3.1}$$

Where $\lambda_0(t)$ is an arbitrary baseline hazard rate and β is regression parameter for treatment, has been estimated based on partial likelihood rather than a full likelihood approach, where $\lambda_0(t)$ is treated as a nuisance parameter function.

Patients used in this study were not identical in terms of the considered risk factors- pain score and PSA level. In this study, effect of these two covariates have been considered through a latent factor - group. Treatment effect was estimated using Cox model after adjusting the effect of group as

$$\lambda(t|x) = \lambda_0(t)\exp(\beta_1 \text{ treatment} + \beta_2 \text{ group} + \beta_3 \text{ treatment}*\text{group}) \tag{3.2}$$

Survival curve based on the Product-Limit estimator of the survival function, proposed by Kaplan and Meier (1958) in equation (3.3) was used to observe the survival pattern of the patients receiving new and standard treatments, among the two latent groups.

$$\hat{S(t)} = \begin{cases} 1 & \text{if } t \leq t_i \\ \prod_{t_i \leq t}[1 - \frac{d_i}{Y_i}] & \text{if } t_i \leq t \end{cases} \tag{3.3}$$

3.2 Latent Group Analysis

A dichotomous group variable was created as a function of pain score and PSA level; with value 1 indicating the asymptomatic group and value 0 indicating symptomatic group, as

$$\text{Group} = \begin{cases} 1, & \text{if pain =low and PSA} < \text{threshold} \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

Expression (3.4) is indicating that, formation of the dichotomous group variable depended on a hidden threshold value of PSA; denoted by *tau* in this study. The next step was to find out an effective threshold value of PSA level.

3.3 Estimation of optimal cutoff point

In this section, we focus on different methods to estimate the cutoff point to create a dichotomous explanatory variable, group as in equation (3.4). Dichotomizing a continuous variable, initiated by Thomsen (1988) [6] makes the model more interpretable. The risk measuring variables- pain score and PSA level has been dichotomized depending on the cutoff point for PSA. We applied different approaches to find out the optimum cutoff point from a set of possible cutoff points using grid search technique as is describing below.

The optimum cutoff point of PSA (*tau*) was calculated in different steps:

1. Grid search method was applied within the inner 70% distribution of the PSA values and

each integer point within this interval was considered as the possible candidate cutoff. We have considered only inner 70% range to avoid having small number of patients in one of the groups following dichotomization.

2. At each candidate cutoff point, dichotomous group covariate was calculated and Cox PH model as in equation (3.2) was fitted

3. Optimum cutoff point was calculated based on the following outcome objects:

   (a) Maximum treatment effect on subgroup

   (b) Maximum goodness of fit component

      i. Maximum likelihood

      ii. Supreme c-index

      iii. Supreme concordance probability estimate (cpe)

3.3.1 Maximum treatment effect on subgroup

As the first approach, the optimum threshold for creating group variable was estimated based on the maximum difference of treatment effect between the asymptomatic and symptomatic patients. Mathematically, we have selected the PSA value which has given maximum interaction effect between treatment and group (in absolute quantity); β_3 in linear predictor of Cox model in equation (3.2). Such that, patients in asymptomatic group were benefited most from the new treatment. At the marginal portion of the considered range, standard error was higher cause of fewer number of patients in one of the subgroups and could estimate unstable subgroup treatment effect. Finally, we considered the best cutoff which provide maximum standardized interaction effect, i.e., $\frac{\beta_3}{SE(\beta_3)}$; or maximum test statistic (or minimum p-value) for the interaction effect parameter β_3.

3.3.2 Maximum goodness of fit components

At the first approach, we have estimated the threshold as the PSA value where the differences of treatment effect was maximum between two groups. But this model may not be the best fitted model. Therefore, we propose some other techniques, based on goodness of fit to estimate the best cutoff.

Maximum likelihood The best cutoff was estimated as the PSA value at which the Cox model (3.2) had maximum likelihood. Log-likelihood was considered in the calculation instead of likelihood.

Supreme c-index C-index was used to estimate the optimum cutoff for PSA and to create group variable. The PSA value which calculated maximum value of c-index for equation (3.2) was taken as the best cutoff.

Harrell et. al. (1982, 1984) [7, 8] proposed the c-index as a way of estimating the concordance probability for survival data is defined as

$$c = \frac{\text{number of concordance pairs}}{\text{number of usable pairs}}$$

Number of usable pairs is computed by counting all pairs $\{(t_i, x_i, \delta_i), (t_j, x_j, \delta_j)\}$ of the observed data, where the smaller follow-up time is an event time. And concordance pairs are the patients pairs, in which the predicted survival times and the observed follow-up times are concordant. If predicted survival times are equal for a patient pair, $\frac{1}{2}$ rather than 1 is added to the count of concordant pairs. The c index measures predictive information derived from a set of explanatory variables in a model, related to rank correlation between observed and predicted outcomes. The value of c-index=0.5 indicates no predictive discrimination (equivalent to Somers' $D = 0$, indicating no correlation) and 1 represents perfect discrimination of patients (Somers' $D = 1$, represents

perfect correlation). Somers' D rank correlation index is related to c-index as $2(c - 0.5)$. In this analysis we used R-package $Hmisc$ to calculate c-index.

Supreme concordance probability estimate As the third component of goodness of fit, we used concordance probability estimate (cpe), another format of concordance probability measure within the frame work of Cox model. PSA value which measured maximum cpe for the Cox model (3.2) was considered as the cutoff point.

Gonen and Heller (2005) [9] estimated c index as a simple function of Cox model, which is more robust to the rate of censoring. They define concordance probability

$$K(\beta) = \text{pr}(T_2 > T_1 | \beta^T x_1 \geq \beta^T x_2)$$

and the concordance probability estimate (CPE) was

$$K_n(\hat{\beta}) = \frac{2}{n(n-1)} \sum \sum_{i<j} \left\{ \frac{I(\hat{\beta}^T x_{ji} < 0)}{1 + exp(\hat{\beta}^T x_{ji})} + \frac{I(\hat{\beta}^T x_{ij} < 0)}{1 + exp(\hat{\beta}^T x_{ij})} \right\}$$

where, $x_{ij} = x_i - x_j$

We used R-package CPE to calculate cpe values.

3.4 Model formulation

Based on different optimization components used in estimating the best cutoff point, we propose four models in two groups in this section. All these models are considered as Cox model with the same linear predictor in equation (3.2), only the latent group variable was defined by different cutoff. In later section we propose another model in Bayesian frame work.

1. Model I: threshold was estimated by maximizing the interaction effect.

2. Model II: threshold was estimated by maximizing the goodness of fit components

(a) Model II(a)threshold was estimated by maximizing the log-likelihood

(b) Model II(b): threshold was estimated by maximizing c-index

(c) Model II(c): threshold was estimated by maximizing concordance probability estimate (cpe)

3. Model III: threshold was estimated using Bayesian change point model (discuss in chapter six)

We checked the proportional hazard assumption for each model.

3.5 Model validation

In this study we estimate the optimum cutoff point and then estimate the model parameters. To check whether the cutoff estimation works perfect, the most stringent test is an external validation- the application of the model to a new population. We have used three-fold cross-validation to estimate the performance of estimated cutoff point. The process flow chart is describing in figure 3

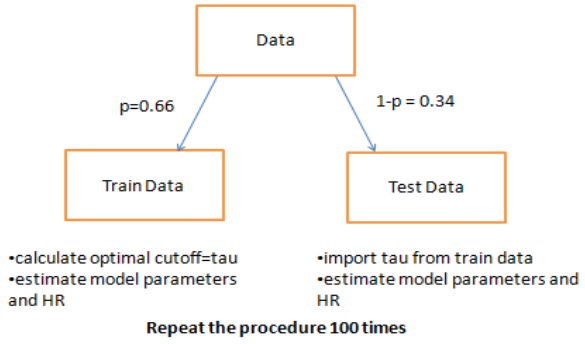

Figure 3: Three-fold cross validation flow chart

Cross-validation is repeated data-splitting. We did the modelling process 100 times, leaving one-third of the data set for testing called test set and estimate the cutoff point using the

techniques in section 3.3 based on remaining two-third objects, train set.

Using the estimated cutoff point, the latent covariate group was created and estimated the regression parameters, and hazard ratio for the train data set. The cutoff points calculated from the train set were applied to the test data set to formulate group variable and fit the Cox regression model (3.2).

R language software version 3.0.1 and also Statistical Analysis System (SAS) version 9.3 have been used in the analysis.

4 Application to the Data

In survival analysis when the interest is to assess the impact of covariates on survival time, Cox proportional hazard model is a popular method to use. The result of Cox model to check only the impact of treatment on survival was significant (p-value=0.019) and the estimated hazard of death for the patients receiving the new treatment was 20.30% lower compared to the patients treated with the standard treatment.

4.1 Estimation of optimum cutoff point

A latent covariate; group was formed as a function of the observed covariates pain score and PSA level. As mentioned in the earlier section the latent covariate depends on hidden threshold of PSA level. Figure 4 indicates the estimate of PSA based on the differences of hazard ratios (model

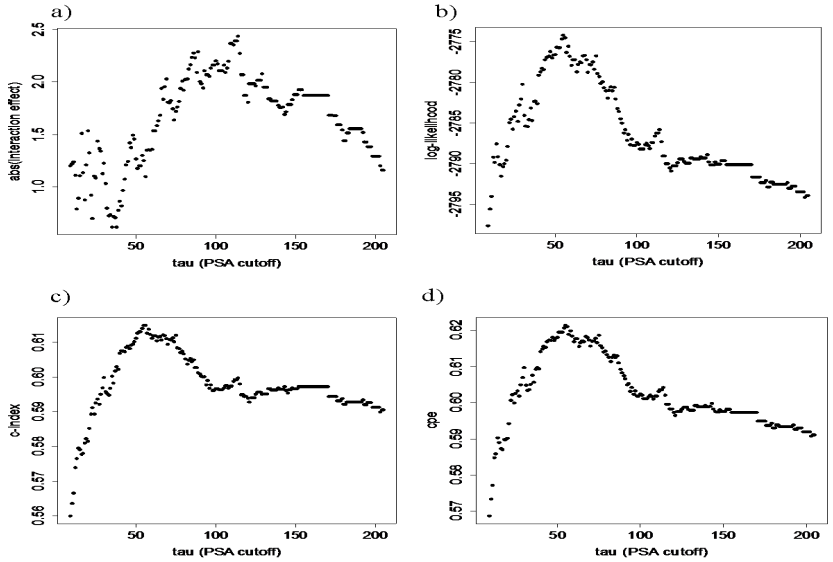

Figure 4: Panel a, b, c and d indicates PSA cutoff based on maximizing the components: a) treatment effect on subgroup, b)log-likelihood, c)c-index, d) cpe

I) was much different from the threshold estimated using the goodness of fit measures. Using

Table 2: Five candidate cutoff points

Tau	Interaction effect	Tau	log-likelihood	Tau	cpe	Tau	c-index
114	-2.4389	55	-2774.17	55	0.6215	56	0.6148
112	-2.3932	56	-2774.49	56	0.6212	55	0.6148
113	-2.3932	54	-2774.67	54	0.6207	54	0.6140
109	-2.3691	58	-2775.57	58	0.6200	58	0.6137
110	-2.3691	51	-2775.65	52	0.6196	52	0.6130

hazard ratio or regression parameter for the interaction effect the cutoff point was estimated at PSA=114, while maximum log-likelihood and maximum cpe provide the same estimate, PSA=55, and another goodness of fit component provide very close estimate PSA=56 as shown in table 2

4.2 Latent group analysis

Based on the estimated cutoff points, latent covariate group was formed and fitted the Cox model as in equation (3.2) The KM survival curve for the two groups of patients based on the cutoff at PSA=114 and PSA=55 is presented in figure 5 In Figure 5(a), differences of the survival probability

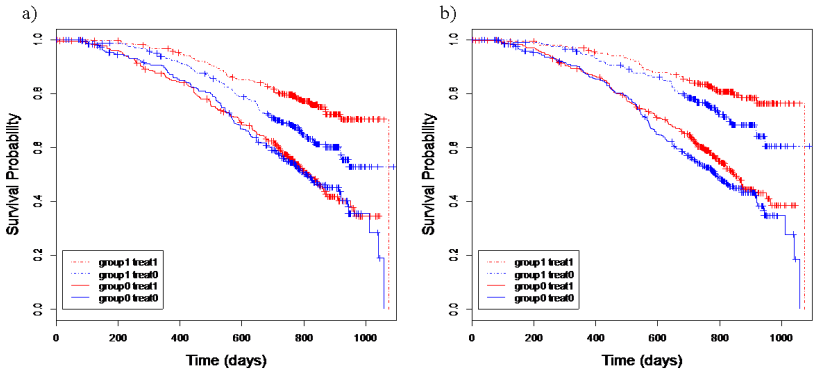

Figure 5: Panel a and b indicates KM survival curves with cutoff at a) maximum treatment effect on subgroup (PSA=114), b) maximum likelihood (PSA=55)

between two treatment groups was much higher for the asymptomatic patients. And the estimated

survival probability pattern was almost similar for two treatment groups in asymptomatic groups; while group was created based on the cutoff at PSA=114. The difference of the survival probability was much lower between two treatments as in figure 5(b) in asymptomatic patients, and also the symptomatic group is showing little difference in treatment effect while the cutff was estimated at PSA=55.

4.2.1 Model I: Maximize subgroup treatment effect

As mentioned in the previous section, threshold was estimated at PSA=114 while maximize treatment effect in subgroup. We defined model I as the Cox model in equation (3.2), with group variable created as following:

$$\text{Group} = \begin{cases} 1, & \text{if pain=low and PSA} < 114 \\ 0, & \text{otherwise} \end{cases}$$

Parameter estimates of model I, is presented in table 3 and the hazard ratio estimates for two groups of patients are in table 4. Negative estimate of main effect of treatment in table 3 indicates the new treatment decreases hazard of death compared to the standard treatment for symptomatic patients, but the difference was not statistically significant. Also the negative group effect describes lower hazard of death for asymptomatic patients compared to symptomatic patients who were treated with the standard treatment. Significant interaction effect indicates the performance of the treatments was different in asymptomatic and symptomatic patients. -2logL

Table 3: Parameter estimates for model I

Parameter	Estimate	SE	p-value
Treatment (new)	-0.0245	0.1234	0.842
Group (asymptomatic)	-0.4923	0.1326	0.0002
Treatment∗Group	-0.487	0.1996	0.0147

for model I was 5571.60 and the AIC was reported as 5577.60. Table 4 is showing performance

Table 4: Hazard ratio estimates for model I

Parameter	Estimate	SE	Lower limit	upper limit	p-value
HR in group 0	0.975	0.120	0.7661	1.242	0.842
HR in group 1	0.599	0.094	0.440	0.815	0.001

of the new drug among two groups; the new treatment has significantly reduced the hazard (by 40.05%) of death among asymptomatic patients. While among the symptomatic patients, the estimated of two treatments were almost similar. This result is expected, as the latent group variable was formed so that the asymptomatic patients would be benefited more with the new treatment, and also supported by the figure 5(a).

4.2.2 Model II: Maximize goodness of fit components

Figure 4 indicates the similar performance of the candidate cutoff points while goodness of fit components (log-likelihood, c-index and and concordance probability estimate) were considering. Table 2 confirms likelihood and cpe provide similar estimate of cutoff point at PSA=55, and c-index results almost similar estimate of best cutoff point at PSA=56.

Model II(a), II(b) and II(c) were defined based on the cutoff points estimated from goodness of fit components. Since model II(a) and II(c) have the same estimate of cutoff, so the explanatory variable group in model II(a) and (c) was defined together as

$$\text{Group} = \begin{cases} 1, & \text{if pain=low and PSA} < 55 \\ 0, & \text{otherwise} \end{cases}$$

While model II(b) considered the estimate of cutoff depending on c-index. The covariate group is

defined as

$$\text{Group} = \begin{cases} 1, & \text{if pain=low and PSA} < 56 \\ 0, & \text{otherwise} \end{cases}$$

Results of model II(a) and (c), and also the result of model II(b) have been presented in table 5. Negative value of the estimated treatment (main) effect indicates new drug reduced the

Table 5: Regression Parameter Estimates of Model II

Parameters	Model II(a), (c)(cutoff = 55)			Model II(b) (cutoff = 56)		
	Estimate	SE	P-value	Estimate	SE	P-value
Treatment	-0.16031	0.10995	0.1448	-0.16904	0.11013	0.1248
Group	-0.83352	0.15124	0.0001	-0.84506	0.15124	0.0001
Treatment*Group	-0.28891	0.23124	0.2115	-0.25301	0.23006	0.2714

hazard of death among the symptomatic patients, but this effect was not statistically significant in all of these three cases. The estimated group effect indicates that asymptomatic patients had ($\exp(-0.83352) = 0.4345$) 56.54% lower hazard of death as compared to symptomatic patients, while treated with the standard treatment. Interaction effect of treatment and group was not statistically significant, indicates no statistical evidence of differences in treatment effect among symptomatic and asymptomatic patients. -2logL for model II(a) and model II(c) was 5548.35, AIC was reported 5554.35, while -2logL for model II(b) was 5548.98 and AIC was reported as 5554.98.

Latent group variable was defined based on different estimates of threshold. And the created latent variable results differences in the estimates of regression parameters and HR as shown in table 3 and table 5. All the goodness of fit components gave almost same pattern of threshold, as in figure 4 and in the next part of the analysis we considered the model II as Cox model defining by cutoff at PSA=55, representing the estimate by maximizing goodness of fit

components. The notable point here is that between two approaches or two models; model I provide significant differences in treatment effect between asymptomatic and symptomatic patients. The other approach, model II do not provide sufficient evidence of treatment differences between symptomatic and asymptomatic patients.

Since both the threshold and hazard ratio were estimated from the same data set, we need to check whether these estimation of threshold is equally good in estimating the model parameters for a new data set. We performed three-fold cross validation to check whether these threshold selection procedures give equally good result when patient variation is taken in to account.

4.3 Cross-validation

100 repeats of 3-fold cross-validation was performed in this study as mentioned in the method section. The interaction effect between treatment and group was overestimated (in quantity) in

Table 6: Cross validation estimates of Model I and II

Parameters	Model I(cutoff =114)			Model II (cutoff = 55)		
	Estimate	train	test	Estimate	train	test
Treatment	-0.02 (0.12)	-0.03(0.09)	-0.14(0.16)	-0.16(0.11)	-0.15(0.07)	-0.16(0.14)
Group	-0.49(0.13)	-0.52(0.14)	-0.69(0.34)	-0.83(0.15)	-0.84(0.13)	-0.75(0.23)
Treat∗Group	-0.48(0.20)	-0.61(0.22)	-0.33(0.39)	-0.29 (0.23)	-0.32(0.18)	-0.32(0.32)
cpe		0.607	0.605		0.622	0.615

the train data set as compared to test data for model I; when the threshold value was estimated at PSA=114 as in table 6. In case of model II, i.e., when the threshold was estimated as PSA=55 the interaction effect also overestimated (in quantity), but for both train and test set and also the rate of overestimation was not so remarkable. The dot plot in figure 6 indicates large variability of the parameter estimates in the test data sets as compared to train data set. Which is expected as smaller number observations were used in the test data set as compared to the train set. Figure

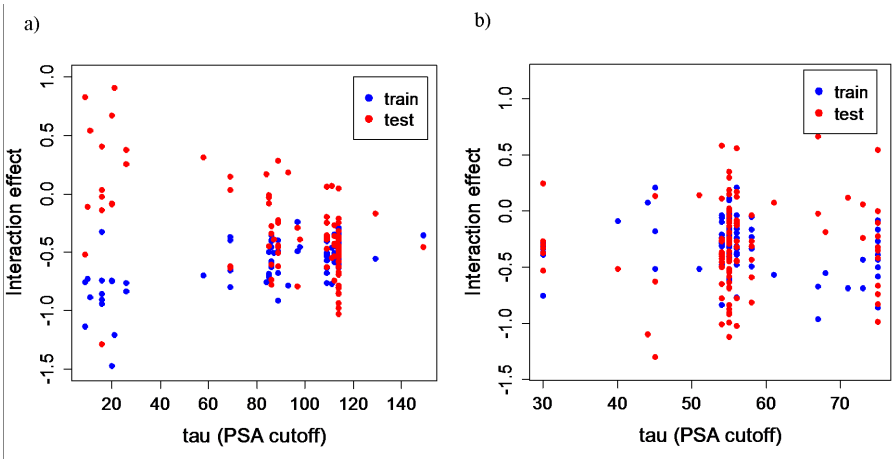

Figure 6: Treatment and group interaction effect estimate in cross validation; panel a) for model I and panel b) for model II

6(a) is presenting the cross-validation estimate of the interaction effect between treatment and group, in train and test data set for model I. This figure also shows the train set estimates had large negative value compared to the test set estimation. Figure 6(b) is presenting the estimates for model II. In some points train estimate was overestimated and also test estimates in some other points Box plot in figure 7(a), supports over-estimation of the hazard ratio for asymptomatic

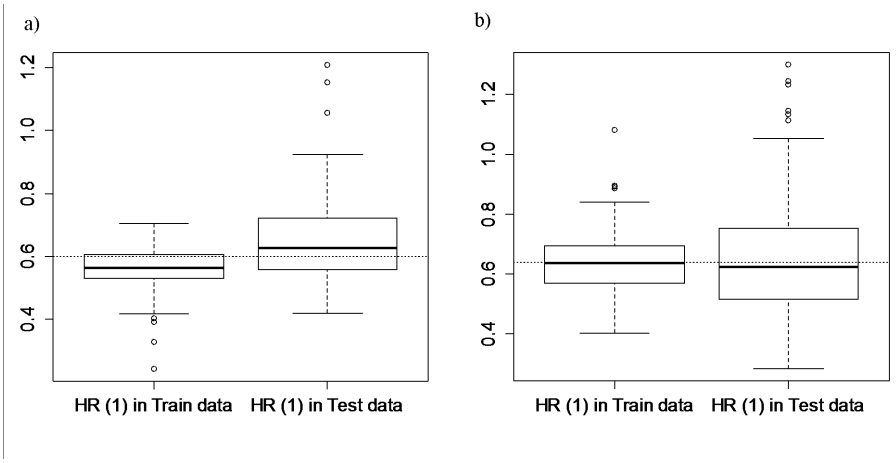

Figure 7: Hazard ratio estimates in cross validation for asymptomatic group; panel a) for model I and panel b) for model II

patients in train data set for model I. Which indicates the selected thresholds do not provide the same good results for new data sets. Figure 7(b) is presenting the cross-validation estimates for asymptomatic patients in model II. Boxplot result for model II does not indicate remarkable over or under estimation in train and test set estimation. The dotted line indicates the estimates of hazard ratio from data set for two models. Estimates in train and test set are more close to the true estimates for model II, compared to model I. Concordance probability estimates (CPE) as in table 6, indicates better discriminatory power of the proportional hazard model represented by model II compared to model I in both train and test set.

5 Simulation

We performed a small simulation study to estimate the treatments effect with making adjustment for the effect of the latent covariate group, also to compare the estimates in train and test set. This simulated data set consisted of (t_i, δ_i, x_i), where t_i was the survival time, δ_i was survival indicator; event($\delta = 1$) and censored (δ=0), and x_i indicated the covariates. The covariates considered this simulated data set were treatment and PSA level. Treatment was considered as a binomial variable having values 1 = experimental treatment and 0=standard treatment with p = Prob(receiving new treatment) = 0.4967. The other covariate PSA level was generated from Uniform distribution within the range $[6, 285]$. Survival indicator (δ) (1=event and 0=censored) was generated as binomial random variable with p = Prob(event=0.3981), similar to the original data.

A proportional hazard relationship was generated from the exponential regression model

$$t_i = \exp(\beta_1 \text{treatment}_i + \beta_2 \text{group}_i + \beta_3 \text{treatment}_i * \text{group}_i)\epsilon_i \tag{5.1}$$

where $\beta_1 = -0.0245, \beta_2 = -0.4923$ and $\beta_3 = -0.487$. The ϵ_i were independent identically distributed exponential variable with rate parameter =6.1. So that the response - survival time (t) was generated as Exponential distribution with rate=$(\eta - 6.1)$. while,

$$\eta = \exp(\beta_1 \text{ treatment} + \beta_2 \text{ group} + \beta_3 \text{ treatment*group}) \tag{5.2}$$

and variable group was created as a function of PSA level as

$$\text{Group} = \begin{cases} 1, & \text{if PSA} < 114 \\ 0, & \text{otherwise} \end{cases}$$

and mean survival time was almost similar to the mean survival time calculated in the main data set. 100 data sets were generated each with 1085 data points.

Cox (1972) model was fitted with linear predictor η as in equation (5.2), and instead of fixing the threshold at PSA=114, new threshold was calculated from the generated data sets. To find out the optimal PSA cutpoint, grid search method was applied on a smaller range of PSA $[80, 130]$. And the similar technique was applied as in the main data set to find out the optimal cutoff point, the maximum interaction effect between treatment and group (β_3) statistics. Let τ_j be the estimated cut point and $\hat{\beta}_{ij}$ the estimate of regression parameters $\beta_i, i = 1,\ 2,\ 3$ in jth simulation, $j = 1, 2, \ldots, 100$. The local squared bias was estimated for β_i as $\hat{b}_i^2 = \{\hat{\beta}_i - \beta_i\}^2$, with $\hat{\beta}_i = \sum_{j=1}^{100} \frac{\hat{\beta}_{ij}}{100}$, and also estimated the local variance as $\hat{v}_i = \sum_{j=1}^{100} \frac{\{\hat{\beta}_{ij} - \hat{\beta}_i\}^2}{100}$. So, the estimate of local mean square error was $\hat{MSE}_i = \hat{b}_i^2 + \hat{v}_i$.

Three fold cross-validation was performed 50 times within each simulation, i.e., two-third of the data set was selected randomly to estimate the threshold tau(τ), and the model parameters were also estimated from this data set, called train set. Using this estimated cut point tau(τ), regression parameters were estimated from the rest part i.e., test set. This was done 50 times within each simulation. At the jth simulation, the cutoff point was estimated as $\tau_j = \sum_{k=1}^{50} \frac{\tau_{jk}}{50}$ for jth simulation and parameters estimated from train set as $\hat{\beta}(tr)_{ij} = \sum_{k=1}^{50} \frac{\hat{\beta}(tr)_{ijk}}{50}$, where $\hat{\beta}(tr)_{ijk}$ was estimate of $\beta_i,\ (i = 1,\ 2,\ 3)$ from the jth simulation in kth train set in cross validation. And the estimate of parameters from train set was $\hat{\beta}(tr)_i = \sum_{i=1}^{100} \frac{\hat{\beta}(tr)_{ij}}{100}$. The local squared bias estimated for β_i from train set was $\hat{b}(tr)_i^2 = \{\hat{\beta}(tr)_i - \beta_i\}^2$, and the estimated local variance was $\hat{v}(tr)_i = \sum_{j=1}^{100} \frac{\{\hat{\beta}(tr)_{ij} - \hat{\beta}(tr)_i\}^2}{100}$. So, the local mean square error was estimated as, $\hat{MSE}(tr)_i = \hat{b}(tr)_i^2 + \hat{v}(tr)_i$.

Parameters estimated at jth simulation, from test set as $\hat{\beta}(te)_{ij} = \sum_{k=1}^{50} \frac{\hat{\beta}(te)_{ijk}}{50}$, where $\hat{\beta}(te)_{ijk}$ was estimate of $\beta_i,\ (i = 1,\ 2,\ 3)$ from the jth simulation in kth test set in cross validation. And the estimate of parameters from test set was $\hat{\beta}(te)_i = \sum_{j=1}^{100} \frac{\hat{\beta}(te)_{ij}}{100}$. The local squared bias estimated for β_i from test set was $\hat{b}(te)_i^2 = \{\hat{\beta}(te)_i - \beta_i\}^2$, and the local variance was $\hat{v}(te)_i =$

$\sum_{j=1}^{100} \frac{\{\hat{\beta}(te)_{ij} - \hat{\beta}(te)_i\}^2}{100}$. So, the local mean square error was $\hat{MSE}(te)_i = \hat{b}(te)_i^2 + \hat{v}(te)_i$.

Different combination of parameters (τ=(100, 114, 130) and β_3=(0.25, 0.487, 0.75)) applied to this data set and estimate the threshold by maximizing the differences of treatment effect between two groups of patients . Further, another parameter combinations ($\tau = 55, \beta_1 = 0.150, \beta_2 = -0.833 \text{and} \beta_3 = (-0.288, -0.50)$ was considered for estimating the threshold by maximizing likelihood (model II). Simulation result (in table 6) indicates that the train set overestimate (in absolute quantity) the interaction effect (β_3) for the first set up (model I). and the test set underestimate (in absolute quantity) the effect. For the second set up (model II) the estimate in train and test set were not much different. Estimates of local bias, variance, mean squared error and relative bias for one parameters combination ($tau = 114, \beta_3 = -0.487$) in first set up, has been shown in table 9 in the appendix.

Table 7: Estimates of treatment and group interaction effect from simulation

Tau	True Value	Without CV	Train	Test
100	-0.25	-0.287	-0.405	0.246
	-0.487	-0.480	-0.582	-0.431
	-0.75	-0.759	-0.840	-0.686
114	-0.25	-0.274	-0.391	-0.241
	-0.487	-0.501	-0.608	-0.461
	-0.75	-0.712	-0.792	-0.6527
130	-0.25	-0.268	-0.390	-0.225
	-0.487	-0.480	-0.559	-0.408
	-0.75	-0.752	-0.805	-0.684
55	-0.288	-0.334	-0.346	-0.330
	-0.50	-0.531	-0.548	-0.506

6 Bayesian Change Point Model

We modeled the time-to-event data using proportional hazard model. Cox's method does not assume any particular distribution for survival times, it only assumes that the effect of the different variables on survival are constant over time and are additive in exponential scale. Partial likelihood (Breslow) likelihood for the considered Cox model

$$L(\beta) = \prod_{i=1}^{n} \left[\prod_{j=1}^{d_i} \frac{\exp(\eta)}{\sum_{l \in R_i} \exp(\eta)} \right]^{v_i} \tag{6.1}$$

where,

$$\eta = \beta_1 \text{treatment} + \beta_2 \text{group} + \beta_3 \text{treatment} * \text{group}$$

β is the parameter vector, n is the total number of observations in the data set, t_i is the ith time which can be either event time or censored time, d_i is the number of failures at time t_i, v_i is censoring indicator (1=event and 0=censored), R_i is risk set at ith event time (t_i), and group is a latent covariate, created depending on the estimate of hidden cutoff point *cp*.

In literature different approaches has been used to estimate hidden value of *cp* [10, 11], such as partial MLE. We have applied the Bayesian approach, where *cp* was assumed to have a distribution instead of a certain value. Estimation of *cp* in Bayesian hierarchical method was done in two steps; apply Bayesian method for Cox model and then use change point technique in Bayesian Cox model.

6.1 Bayesian Cox Model

In Bayesian frame, work regression parameters of Cox model are considered to have some distribution, instead of having a fixed value. The variable *group* has been created in this section based on the estimated threshold in model I and model II. Vague normal prior distribution $N(0, 10^6)$ were

considered for the model parameters (βs). Posterior distribution of the regression parameters, βs are

$$\pi(\beta|D) \propto L_p(D|\beta)p(\beta) \tag{6.2}$$

where $L_p(D|\beta)$ is the likelihood function with regression coefficients (β) as parameters. Formulation of marginal posterior distribution is not straight forward, as the partial likelihood of Cox model does not assume any distribution. $Markovchain Monte Carlo$ technique has been used to solve this problem. We have used SAS (version 9.3) procedure Proc MCMC based on blocked Metropolis(-Hastings) algorithm. In Proc MCMC, it is required to specify the likelihood function and the prior as mentioned in equation (6.2). Construction of likelihood function is not straight forward in Cox model, as no certain distribution is assumed for the response, survival time. As seen in equation (6.1) the likelihood depends on the risk set at each event time. The risk set consists of all data point j such that $j \leq i$, when the time variable is ordered in descending order. The data set considered in this study have multiple events at some event time point. So, it was needed to check whether subsequent observations i.e., observations i, $i+1$, $i+2$ have the same survival time as t_i. If more than one observations have the same event time, all tied observations need to be included in the risk set calculation. That is, in calculating likelihood function it is required to access both the previous and the subsequent observations in the data set. In Proc MCMC, we have used LAG function to do it, and then cumulatively increment in the survival function in SAS statements to create the likelihood function.

6.2 Change point model

In this section we have considered latent covariate, group correspond to a change point cp, where

$$group = \begin{cases} 1, & \text{if pain score=low and PSA} < \text{cp} \\ 0, & \text{elsewhere} \end{cases}$$

Which indicates hazard ratios and also the mean survival times are different for the patients having lower and higher level of PSA compared the change point.Along with the regression parameters, the change point was considered as a random variable. Prior for the change point was considered as Uniform distribution over the range (9,205), the inner 70% distribution of PSA level in the data set. Then the joint distribution of likelihood and prior was

$$\pi(\beta, cp|D) \propto L_p(D|\beta, cp)\ p(\beta)\ p(cp) \tag{6.3}$$

Marginal posterior distribution of cp, i.e., the distribution of $cp|Y$ requires integration over βs, which also required formation of joint distribution. This has been done through *Markov chain Monte Carlo* sampling using Proc MCMC in SAS 9.3 version. Results for the

Table 8: Posterior summary measures of the parameters

Parameter	Mean	SD	MCSE	2.5% HPD	97.5%	MCSE/SD	Sample
Treatment	-0.158	0.110	0.003	-0.385	0.048	0.030	1000
Group	-0.820	0.154	0.005	-1.148	-0.544	0.033	1000
Treatment$*$ Group	-0.305	0.232	0.006	-0.771	0.098	0.027	1000
cp	55.629	5.72	0.198	43.923	69.300	0.034	1000

Bayesian change point model is presented in table 8. Posterior mean of cp is 55.617 with 95% credible interval (43.26, 69.45). Only main effect of group is significant, which indicates the hazard was lower for asymptomatic patients compared to symptomatic patients. The main and interaction effect of treat included 0 in the 95% credible interval indicates no significance differences of treatment effect among asymptomatic and symptomatic patients. The reported value of DIC was 5559.78, deviance evaluated at the posterior mean, D_{mean} was 5547.82 and the effective number of parameters (P_D) were 5.974.

The posterior autocorrelations in figure A: 9 were very low for all lags. Again the trace

plots in figure A: 10 showed good mixing for all parameters. Further Geweke (test result is in table 11) and Heildelberger-Welch diagnostics indicated good mixing, but the chain was too small for the Raftery-Lewis diagnostic. The MC error were below 5% of the posterior SD. Posterior

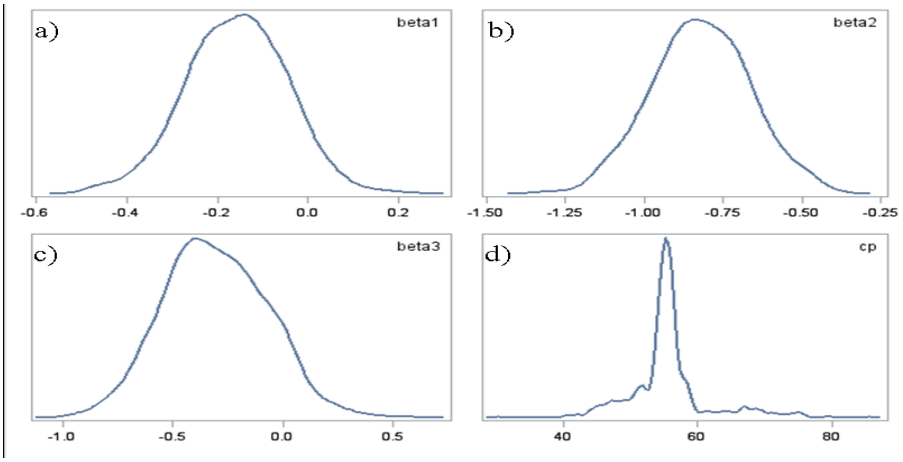

Figure 8: Panel a, b, c and d is showing the posterior density plots of the effects: a)treatment, b)group c) treatment∗group d)change point estimate

density plot (fig 8) indicates normality for the regression parameters β_1, β_2 and β_3, denoting the main effect of treatment, group and their interaction effect. Also the change point estimate, *cp* shows approximate normal distribution pattern.

7 Discussion

The main objective of this study was to find best cutoff point (PSA value) to separate patients into two groups; asymptomatic and symptomatic and estimate the treatment effect. To do this, a latent covariate was formed as a function of two observed covariates; pain score and PSA level and depended on the the threshold value of PSA.

We have applied systematically different approaches to estimate the best cutoff point, and also to estimate the treatment effect. At the first approach, we estimated best cutoff point by maximizing the object component; differences in treatment effect in two groups. Creating group based on this estimate of threshold, provided significant separation of patients in terms of treatment effect. Hazard ratio (of death) for the treatment (new treat vs. standard treat) was much lower among asymptomatic patients as compared to the symptomatic patients. In the second approach, thresholds of PSA was estimated using the best fitted point based on likelihood, Harrells' c-index and concordance probability. All these three estimates were almost the same. And we considered the estimate using likelihood for comparing the results with other approaches. Patients grouped based this estimated threshold, was not significantly different in treatment effect. Which indicates patients' response on treatment was not different among these two groups of patients.

In the third approach, we used the change point model in Bayesian framework for estimating the threshold and the model parameters. The threshold has been considered as the change point and was estimated using both the prior distribution assumptions for the change point and the data set. model used only one model and use the data once to estimate all parameters.

Threshold estimated at the first approach, i.e., model I provide significant differences between the treatment effect in asymptomatic and symptomatic group. But this model did not provide the best fitting to the data (maximum likelihood or minimum AIC). While the second

approach, i.e., model II provide lower AIC compared to model I. Also the first approach failed to provide similar good estimate of model parameters for new data set. While, the second approach provide more stable results in case of fitting the model for new data set.

Threshold estimation method described above used the data set twice; to estimate threshold and then estimate the model parameters. Which could overestimate or underestimate the effects. In this study estimated value of the change point was quite similar to the threshold estimated in the second approach. The change point model provide the shorter confidence interval as compared to frequentist, for estimating the hazard ratio among asymptomatic patients (table 11, in the appendix).

And finally we concluded that the patients receiving the new treatment were expected to have longer survival time compared to the patients treated with the standard one. But while treated with new treatment, the estimated survival time was not different in asymptomatic and symptomatic group.

7.1 Recommendations

Appropriate treatment options usually correspond with risk level. Prostate Cancer Research Institute suggested a general classification of risk groups based on PSA level, PSA density in blood, PSA velocity, size of lump, % biopsy cores positive and Gleason sum. In this study the effectiveness of the treatment were assessed between two groups by using pain score and PSA level only. So, considering other important risk factors could be better options to find cutoff point.

In Bayesian change point model we considered Uniform prior for change point within inner 70% range of PSA. Instead, some informative prior assumption(exponential, weibull etc.) could be considered depending on the distribution of PSA.

References

[1] WIKIPEDIA. Prostate cancer. http://en.wikipedia.org/wiki/Prostate_cancer, August 2013. [Online; accessed 7-August-2013].

[2] Prostate Cancer Coalition of North Carolina (PCCNC). The Prostate Specific Antigen (PSA) Blood Test. http://www.pccnc.org/early_detection/psa/, 2013. [Online; accessed 29-August-2013].

[3] National Cancer Institute: at the National Institutes of Health. Prostate cancer treatment (pdq). http://www.cancer.gov/cancertopics/pdq/treatment/prostate/Patient/page1#Keypoint4. [Online; accessed 7-August-2013].

[4] J. S. Kim. Cutpoint analysis with bootstrap validation for the cox proportional hazards model. http://web.pdx.edu/~jong/WSU_Coll04.pdf, 2004. [Online; accessed 29-july-2013].

[5] B.A. Williams, J. N. Mandrekar, S.J. Mandrekar, S.S. Cha, and A.F. Furth. Finding Optimal Cutpoints for Continuous Covariates with Binary and Time-to-Event Outcomes. Technical report, Department of Health Sciences Research, Division of Biostatistics, Mayo Clinic, Rochester, Minnesota, 2006.

[6] B.L. Thomsen. A note on the modeling of continuous covariates in cox's regression model. Technical report, 88/5, Statistical Research Unit, University of Copenhagen, 1988.

[7] F.E. Harrell, R.M. Califf, D.B. Pryor, K.L. Lee, and R.A. Rosati. Evaluating the yield of medical tests. *J. Am. Med. Assoc.*, 247:2543–6, 1982.

[8] F.E. Harrell, K.L. Lee, R.M. Califf, D.B. Pryor, and R.A. Rosati. Regression modelling strategies for improved prognostic prediction. *Statist. Med.*, 3:143–52, 1984.

[9] M. Gonen and G. Heller. Concordance probability and probability and discriminatory power in proportional hazards regression. *Biometrika*, 92(4):965–970, 2005.

[10] U. Jensen and C. Lutkebohmert. Change-point models. *Encyclopedia of Statistics in Quality and Reliability*, 2013.

[11] D. M. Zucker, S. Agami, and D. Spiegelman. Testing for a changepoint in the cox survival regression model. *Journal of Statistical Theory and Practice*, 7(2):360–380, 2013.

[12] D.R. Cox. Regression model and life tables (with discussion). *J.R. Statist. Soc.*, B 34:187–220, 1972.

[13] F. E. Harrell, K.L. Lee, and D.B. Mark. Tutorial in biostatistics multivariate prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387, 1996.

[14] J. P. Klein and L. M. Melvin. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 1997.

[15] and Lemeshow S. Hosmer Jr, D. W. and S. May. *Applied survival analysis: regression modeling of time to event data*, volume 618. Wiley, 2011.

[16] L. Emmanuel and A. B. Lawson. *Bayesian Biostatistics*. Wiley, 2012.

[17] National Cancer Institute at the National Institute of Health. Prostate-Specific Antigen (PSA) Test. `http://www.cancer.gov/cancertopics/factsheet/detection/PSA`. [Online; accessed 1-September-2013].

A

Graphical plots

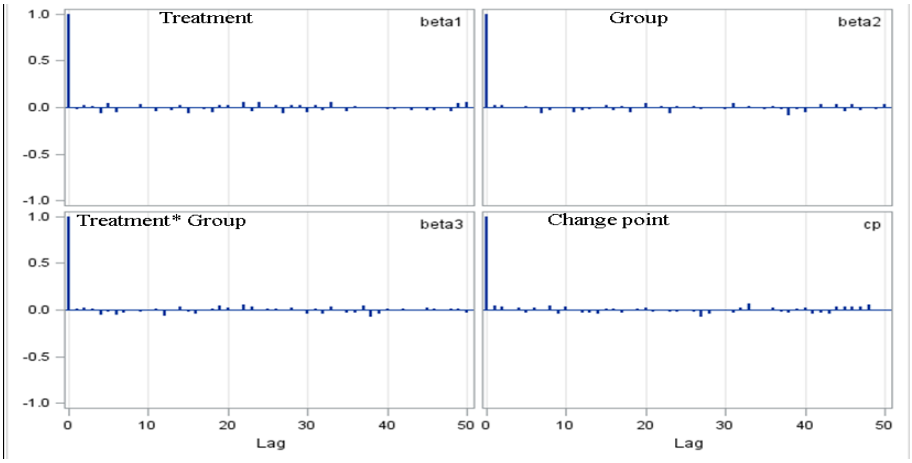

Figure 9: Autocorrelation plot for treatment, group, treatment∗group and change point

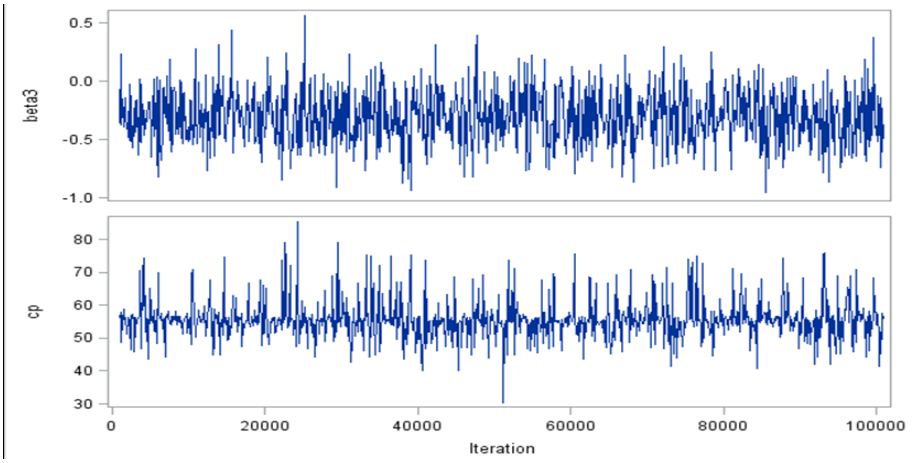

Figure 10: Trace plot for treatment∗group (beta3) and change point (cp)

B

Tables

Table 9: Parameter estimate from simulation

Parameter	True Value	Estimates	Without CV	Train	Test
Treatment	-0.0245	Mean	-0.0165	0.0128	-0.0433
		Bias	0.0079	0.0373	0.0188
		Variance	0.0135	0.0132	0.0145
		MSE	0.0135	0.0146	0.0149
		Relative bias	32.55%	152.50%	-77.11%
Group	-0.4923	Mean	-0.4866	-0.3681	-0.4348
		Bias	0.0056	-0.3681	0.0574
		Variance	0.0182	0.0145	0.0164
		MSE	0.0183	0.0299	0.0197
		Relative bias	1.14%	25.21%	11.67%
Treatment∗Group	-0.487	mean	-0.5011	-0.6088	-0.4616
		Bias	0.0141	0.1218	0.0253
		Variance	0.0377	0.0331	0.0387
		MSE	0.0379	0.0480	0.0393
		Relative bias	-2.90%	-25.01%	5.21%
MCE				0.0437	0.0438
cpe			0.5954	0.5876	0.5889
c-index			0.5958	0.5894	0.5905
Tau	114				110.01

Table 10: Geweke diagnostics

Parameter	z	Pr $> \|z\|$
Treatment	-0.685	0.493
Group	-0.222	0.823
Treatment * Group	0.252	0.800
Change point	0.362	0.716

Table 11: Hazard ratio estimates for asymptomatic patients

Tau	Estimate	SE	lower limit	upper limit	interval
55	0.6381	0.1298	0.4283	0.9507	0.5224
56	0.6557	0.1324	0.4414	0.9742	0.5328
114	0.5995	0.0941	0.4408	0.8155	0.3747
55.629	0.6423	0.1306	0.4311	0.9569	0.5258
Posterior estimate	0.6428	0.134	0.427	0.935	0.508

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Analysis of prostate cancer data at the presence of latent factor: Estimation of the treatment effect

Richting: **Master of Statistics-Biostatistics**
Jaar: **2013**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Polin, Afroza

Datum: **11/09/2013**