Subgroup Analysis of Censored Data on Cancer Treatment

Qing Zhang[1], Hongshik Ahn[2]
Department of Applied Mathematics and Statistics, Stony Brook University,
Stony Brook, NY 11794-3600

Abstract
In this study, we develop a statistical method based on tree-structured classification to assign patients from different treatment groups into subgroups with different medical recommendations. Since it is difficult to discover treatments that benefit all patients, we want to identify subgroups of patients for whom the treatment has an enhanced effect. Our model is applied to survival data. We classify terminal nodes into subgroups by comparing the relative event rates of the terminal nodes and the relative event rates of their corresponding immediate predecessors. Given the suggested subgroup of each terminal node, we give a suggested method on how to identify which of the splitting variables are keen to which of the subgroups by tracing back along the tree. Performance of our proposed method is evaluated from multiple simulation runs. The result shows that our method is more likely to give an appropriate recommendation when the treatment effect is more heterogeneous.

Key Words: Censoring, Proportional hazards regression, Recursive partitioning, Subgroup analysis, Survival analysis.

1. Introduction

1.1 Background

Subgroup analysis is a way of finding out if the treatment effect or other intervention was more effective for some people than others, and extract as much information as possible from the available data. General approaches for subgroup analysis can be either pre-planned, or start with a post-hoc manner. If there exists a prior hypothesis of the treatment effect being different in certain subgroups, then the evaluation of the hypothesis should be part of the study design (CPMP, 1995). Meanwhile, subgroup analysis is also used in post-hoc manner to explore unusual and unexpected results (Chow and Liu, 2004). Ciampi, Negassa and Lou (1995) introduced a tree-structured subgroup analysis using the RECursive Partitioning and AMalgamation (RECPAM) algorithm. The partitioning is based on some basic characteristics, such as demographics and clinical measurements (Negassa et al., 2005). Su et al. (2009) introduced similar method called "interaction tree" (IT). They introduce the idea of assessing the interactions to determine the split of the tree.

In this paper, we propose a data-driven tree structured classification method to explore the heterogeneity structure effect across subgroups which are defined beforehand. The problem we are focusing on is trying to determine subgroup of patients who will get benefit from certain treatment. Since it is difficult to find a treatment that benefits all patients, identifying subgroups of patients for whom a treatment has an enhanced effect is crucial. For patients with non-small-cell lung cancer (NSCLC), surgery is still the basic treatment method, and the benefit of having Adjuvant Cisplatin-Based Chemotherapy (ACT) was not being fully demonstrated until a decade ago. Chemotherapy has advantages in having high tumor relapse. However, due to its toxicity characteristics, there always exist potential risks for patients to receive chemotherapy. Thus, doctors must weigh between ACT and observation (OBS) to optimize the survival time and life quality for their patients. In this paper, we present our method based on recursive partitioning and regression trees (Breiman et al., 1984). The applications of tree structured classification methods have been developed

since CART (classification and regression trees) has been introduced by Breiman et al. (1984). By recursively bisecting the predictor space, the hierarchical tree structure partitions the data into subgroups and available predictions of the underlying relationship between the response and its corresponding predictors. In our study, the partition of the tree is based on gene expressions, and the final subgroup is determined by comparing the relative event rates of the terminal nodes and its corresponding immediate predecessor. Performance of our proposed method is evaluated using 100 simulation runs.

In the remaining of this paper, we will introduce the method in detail, then followed by the simulation designs and results. Finally, we will apply our method to the real data set, and conclude our analysis by a brief discussion.

1.2 Method

First, we define the relative event rate of node *i* as the proportion of death in node *i* relative to the root of the tree. That is:

$$\frac{N_{death\,(i)}/N_{all\,(i)}}{N_{death\,(1)}/N_{all\,(1)}}$$

Where $N_{death\,(i)}$ is the total number of death subjects in node *i*, and $N_{all\,(i)}$ is the total subjects in node *i*. We let *i* indicates the node number, so the node number of the root is 1.

The method we use to give medical recommendations is as follows:
H_i: Relative event rate of the immediate predecessor of the terminal node where subject *i* falls into.
h_i: Relative event rate of the terminal node where subject *i* falls into.
OBS_i: Number of subjects with OBS treatment in the terminal node where subject *i* falls into.
ACT_i : Number of subjects with ACT treatment in the terminal node where subject *i* falls into.
Treatment recommendation is given according to the following strategy.
If $h_i < H_i$, $OBS_i > ACT_i$: chemo not recommended (chemo not rec)
If $h_i > H_i$, $OBS_i > ACT_i$: chemo recommended (chemo rec)
If $h_i < H_i$, $OBS_i < ACT_i$: chemo recommended
If $h_i > H_i$, $OBS_i < ACT_i$: chemo not recommended
If the initial grown tree only has one single node, then the decision is inconclusive.

2.Simulation

2.1 Simulation Designs

We apply our method to two different simulation designs. The first design has a more heterogeneous treatment effect, and in the second design, the treatment effect is more homogeneous.
Survival times with right censoring are generated from Cox proportional hazards model (Cox, 1972), given as the following formula,

$$t = \left(-\frac{\log(v)}{\lambda \exp(\boldsymbol{X}'\boldsymbol{\beta})}\right)^{\frac{1}{\rho}}$$

where v is a uniform variate on $[0,1]$, λ is the scale parameter and ρ is the shape parameter in Weibull distribution, $\boldsymbol{\beta}$ is the coefficient matrix from Cox proportional hazards model. Censoring times are generated from exponential distribution with 20% and 40% censoring rates. Each dataset consists of 2000 subjects with 103 variables. Variables include: continuous survival time Y in days, censoring indicator (death/alive), binary treatment, covariates $X_1 \sim X_5$ generated from unif $[-1,1]$ are considered common

factors in disease and non-disease patients, covariates $X_6 \sim X_{10}$ genearated from unif $[-1,5]$ are considered as factors that are in favor of chemo therapy, and $X_{96} \sim X_{100}$ generated from unif $[-5,1]$ are considered as factors that are not in favor of chemo therapy. The remaining covariates $X_{11} \sim X_{95}$ generated from unif $[-0.5, 0.5]$ are factors not related to the treatment.
Subjects are generated from each of the four subgroups: ACT treatment and chemo recommended, ACT treatment and chemo not recommended, OBS treatment and chemo recommended, OBS treatment and chemo not recommended. For Design 1, the median survival times for the subgroups mentioned before are: [2.47, 1.23, 0.45, 2.96]. For Design 2, the median survival times for the four subgroups are: [2.47, 1.23, 1.17, 2.43]. The first two subgroups: ACT_rec and ACT_notrec are the same in both designs. On the other hand, survival times for OBS_rec and OBS_notrec in Design 1 and 2 are differ by the rate parameters in the exponential distribution when we are generating survival times.

2.2 Simulation Results

We apply our method to two different simulation designs. The treatment effect was statistically significant using the Cox proportional hazards model for Design 1 [95% CI (0.76, 0.97), p=0.013]. For Design 2, the treatment effect was not statistically significant using the Cox model [95% CI (0.98, 1.23), p=0.106]. So the treatment effect for the first design is more heterogeneous than the second design. Both single and multiple simulation runs were performed. Initial trees are grown using the data set generated from the simulation designs, then the full trees are pruned by using 1 standard error rule.

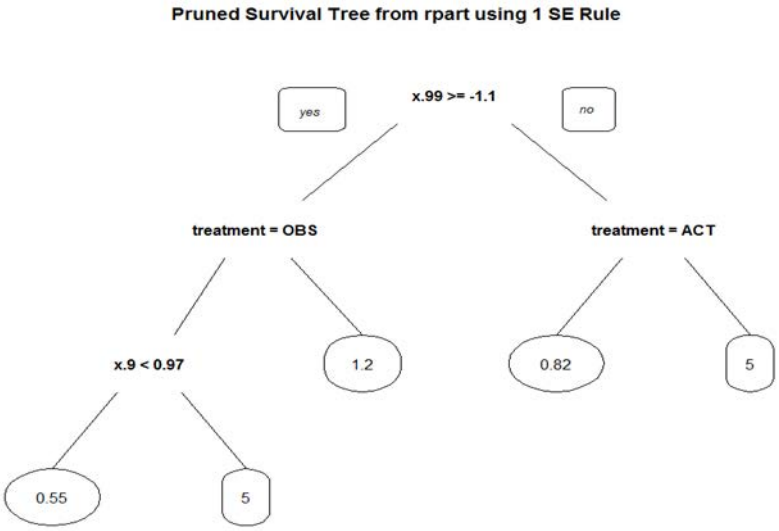

Figure 1: Pruned Tree for Design 1 from a single simulation run, 40% censoring.

Figure 1 shows the rpart tree from Design 1 from a single simulation after 1 standard error pruning, and node numbers are assigned from top to bottom, and left to right. Numbers on the terminal nodes are the relative event rates. Thus, we can make a deduction on the splitting variables as follows:
Node 6: lower relative event rate (0.82) for x99- (x99<-1.1) and ACT. This suggests that x99 is not in favor of chemo.
Node 7: very high relative event rate (5) for x99- and OBS, this supports the prediction that x99 is not in favor of chemo.
Node 8: lower relative event rate (0.55) for x99+, OBS and x9-. This suggests that x9 is in favor of chemo, also supports the prediction that x99 is not in favor of chemo. Node 9: higher relative event rate (5),

comparing it with node 8, the only difference is made by x9: we have x9+ for node 9. This also supports the prediction that x9 is in favor of chemo.
Node 5: high hazard rate (1.25) for x99+ and ACT, suggests that x99 is not in favor of chemo.
Notice that if we compare node 5 with node 6, both nodes are assigned to ACT and the difference is that x99+ is assigned to node 5 and x99- is assigned to node 6. According to the predictions above, chemo given in node 5 is inappropriate and OBS in node 6 is an appropriate treatment. This can explain the higher relative event rate in node 5 than in node 6.

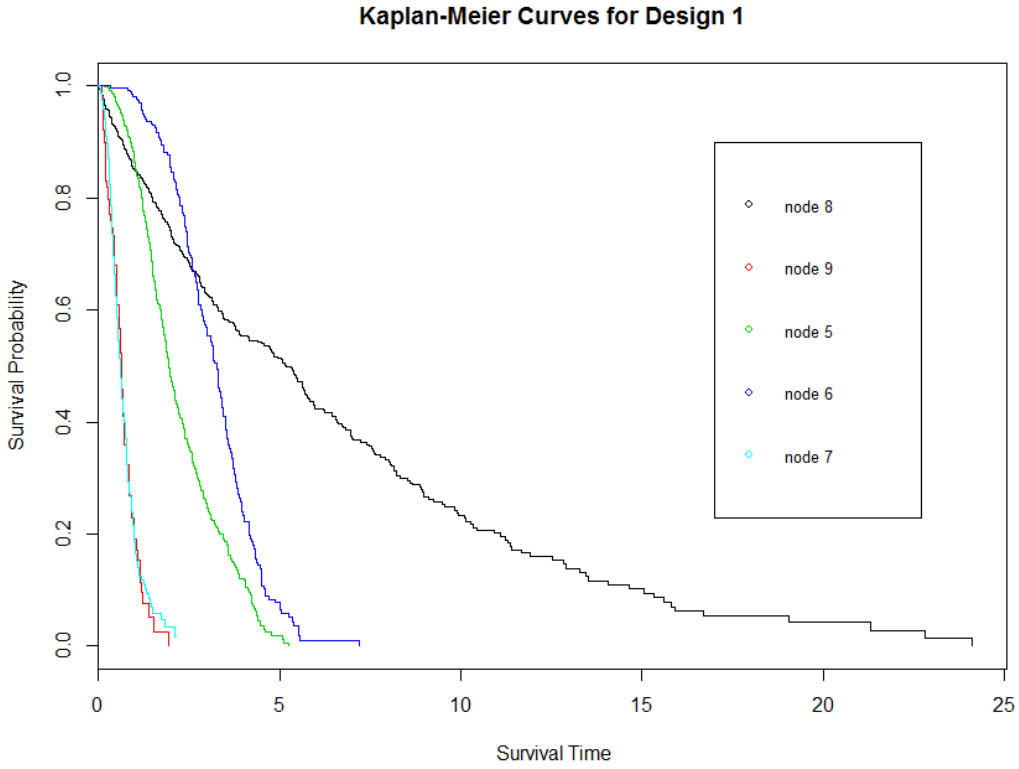

Figure 2: Kaplan-Meier curve for Design 1 from single simulation run, 40% censoring.

Figure 2 shows the Kaplan-Meier curve for a single simulation run from Design 1 with a 40% censoring rate. The curves are well separated, except node 5 and node 7. For multiple simulation runs, we performed 100 times by using the whole data, and got the mean accuracy from each simulation runs. The mean accuracy for Design 1 with 20% censoring was 0.92, and was 0.91 with 40% censoring.

Table 1: Frequency of splitting variables from 100 simulation runs with 40% censoring.

Name	Frequency
Treatment	100
x.97	31
x.96	31
x.98	29
x.7	28
x.9	26
x.100	25
x.99	25

x.8	22
x.6	22
x.10	19

Table 1 lists the most frequently appeared variables in the rpart trees from 100 simulation runs, and it shows that variables with high frequencies are the ones related to classifying chemo or non-chemotherapy.
Same procedures also applied to Design 2. Figure 3 shows the rpart tree with 1SE pruning for Design 2, and Figure 4 illustrates the corresponding Kaplan-Meier Curves for each terminal nodes. The prediction accuracy for Design 2 from this single simulation run was 0.725.

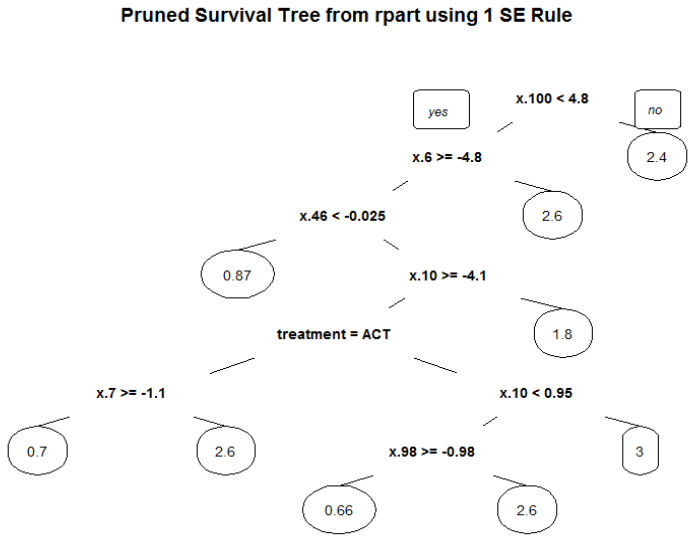

Figure 3: Pruned Tree for Design 2 from a single simulation run, 40% censoring.

As we can see from the figure above, from node 8 and below, the tree structure is similar to what we had before: split in treatment first, and followed by some significant variables along the paths to the leaves. Since the splitting variables are more important as it is more close to the bottom layer of the tree, when we do the verification of the correctness of the simulation, we ignore the part that lie above node 8 if the terminal node has treatment as one of the splitting variable when tracing back from it to the root of the tree. Otherwise, we obtain how many ACT and OBS treatment in each terminal nodes, and also the hazard rates for each of the treatment groups in the terminal nodes.

Node 3 has x100+ and an overall hazard rate of 2.4. As we exam further for this node, we have 17 observations from ACT group, and 11 observations from OBS group. Furthermore, all of the 17 observations in the ACT group are from chemo_notrec class, and all the 11 observations from OBS group are also from chemo_notrec class. The hazard rate for ACT group (4.6) is higher than that for OBS group (0.85). This support the assumption that x100 is not in favor of chemo.

Node 5: 14 observations from ACT group, and 11 observations from OBS group. All of the observations in this node are from chemo_notrec class, with an overall hazard rate of 2.6. The hazard rate for the ACT group was 3.2, and is higher than the hazard rate for OBS group (2.0). Since we have x6- for this terminal node, this suggests that x6 is in favor of chemo.

Node 9: 46 observations from ACT group, and 32 observations from OBS group. All of the observations are from chemo_notrec class, with hazard rate of 2.15 for the ACT group, and 1.4 for the OBS group. Again,

we have higher hazard rate for ACT group than OBS group. We have x10- as the immediate splitting criteria for node 9, and this suggests x10 is in favor of chemo.

Node 6: 462 observations from ACT group and 457 observations from OBS group, with an overall hazard rate of 0.87. The hazard rate for ACT group is 0.85 and the hazard rate for OBS group is 0.89, just a little bit higher than the ACT group. The immediate splitting variable for this terminal node is x46, and suggests that x46 might not be a significant factor for classifying subgroups of chemo and non-chemo, and take a detailed look at the performance of classification on chemo and non-chemo, among ACT group, 238 observations are from chemo rec class, and 224 observations are from chemo notrec class. Among OBS group, 219 observations are from chemo rec class, and 238 are from chemo notrec class. Therefore, the performance of classifying individuals into different classes is not very promising.

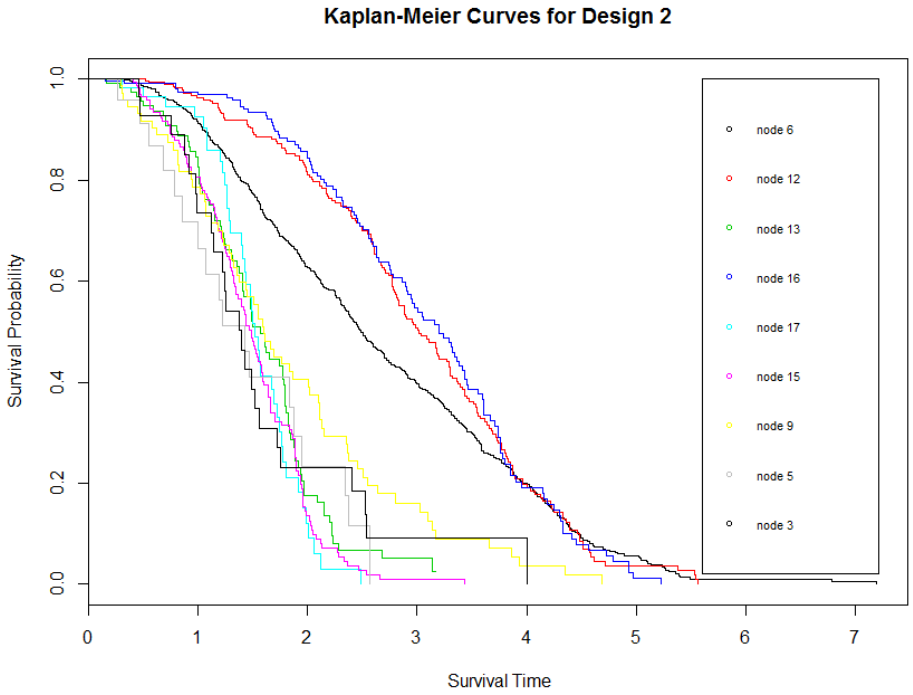

Figure 4: Kaplan-Meier curve for Design 2 from a single simulation run, 40% censoring.

Figure 4 shows the Kaplan-Meier curve for a single simulation run from Design 2 with a 40% censoring rate. The curves are not well separated as Design. We also performed 100 simulation runs by using the whole data, and got the mean accuracies from each simulation runs. The mean accuracy for Design 2 with 20% censoring was 0.8, and was 0.72 with 40% censoring.

Table 2: Frequency of splitting variables from 100 simulation runs with 40% censoring (top 11).

Name	Frequency
treatment	71
x.6	33
x.10	26
x.97	25
x.98	21

x.96	20
x.100	20
x.7	19
x.99	15
x.8	14
x.9	14

3. Real Data Analysis

To illustrate our approach, we apply our methods to two data sets. One is the data from the National Center for Biotechnology Information (NCBI) series 37745: biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis and tissue microarray validation (Botling J, Edlund K, Lohr M, Hellwig B et al., 2013), and the other one is from the retinopathy data set under survival package in R (Blair et al., 1976).

The NCBI data set consists of 196 non-small cell lung cancer (NSCLC) subjects who have received surgical treatment from 1995 through 2005 with 26% censoring. The data set has 16380 variables: time-to-event in days, censoring indicator (death/alive), age (in years), gender, treatment (yes/no), and gene expression variables. Gene expression array data were used as a training set to screen for single genes with prognostic impact (Bolting et al., 2013).

The treatment effect was not statistically significant using the Cox proportional hazards model [95% CI (0.60, 1.60), p=0.93]. However, we are interested in searching for subgroups of patients for whom the alternative treatment would be better than the current treatment. Various folds of cross validation were performed. In the real data set, we know a patient received what kind of treatment, but we do not know the correct treatment the patient should receive. In this case, we can make predictions of the correct treatment based on the method we described in Section 2.2. The accuracies from cross validations are very low because we compared the predictions with the treatment patients received, but not the correct treatment patients should receive. With all the data from Series 37745, Figure 5 illustrates the unpruned rpart tree. Unpruned tree for Design 2 is shown below, because the 1 standard error rule pruning method as we used before resulted in no split. When we compare the prediction for the whole data set with the actual treatment, the prediction accuracy is 0.58.

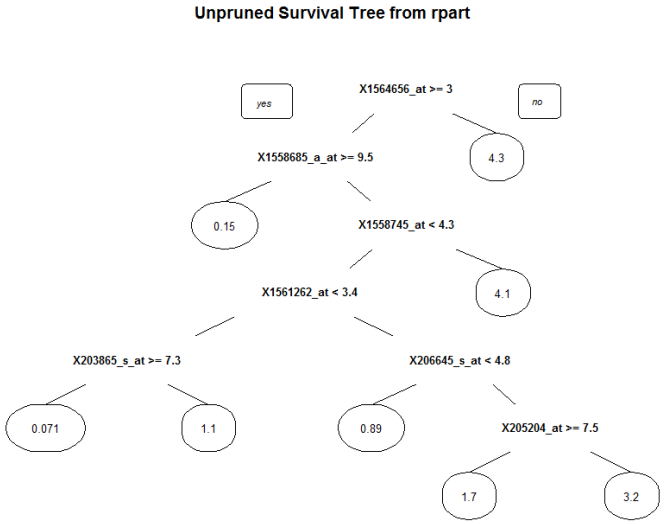

Figure 5: Unpruned tree for Series 37745

Table 3: Illustration of the Split for the Unpruned Tree using NCBI Data Set

Node Number	Number of ACT	Number of OBS	Total	Prediction
3	8	5	13	Chemo notrec
4	12	3	15	Chemo rec
7	6	1	7	Chemo not rec
10	8	2	10	Chemo rec
11	8	6	14	Chemo not rec
12	10	4	14	Chemo rec
14	14	6	20	Chemo rec
15	5	2	7	Chemo not rec

Node 3: $X1564656^{-}$, high relative event rate (4.3), and the number of ACT subjects is greater than OBS. This suggests that X1564656 is in favor of chemo.

Node 4: $X1558685^{+}$, low relative event rate (0.15), and the number of ACT subjects is greater than OBS. This suggests that X1558685 is in favor of chemo. This group has X1564656>=3. Thus, the conclusion for Node 3 is supported.

Node 7: $X1558745^{+}$, high relative event rate (4.1), and the number of ACT subjects is greater than OBS. This suggests that X1558745 is not in favor of chemo. This group has X1558685<9.5. Thus, the conclusion for Node 4 is supported.

Node 10: $X203865^{+}$, low relative event rate (0.071), and the number of ACT subjects is greater than OBS. This suggests that X203865 is in favor of chemo.

Node 11: $X203865^{-}$, high relative event rate (1.1), and the number of ACT subjects is greater than OBS. This suggests that X203865 is in favor of chemo.

Node 12: $X206645^{-}$, low relative event rate (0.89), and the number of ACT subjects is greater than OBS. This suggests that X206645 is not in favor of chemo.

Node 14: $X205204^{+}$, low relative event rate (1.7), and the number of ACT subjects is greater than OBS. This suggests that X205204 is in favor of chemo.

Node 15: $X205204^{-}$, high relative event rate (3.2), and the number of ACT subjects is greater than OBS. This suggests that X205204 is in favor of chemo. The conclusions for Node 14 and Node 15 agree.

Node 8: $X1561262^{-}$, low relative event rate (0.54), and the number of ACT subjects is greater than OBS. This suggests that X1561262 is not in favor of chemo.

Classification of chemo nodes: 4, 10, 12, 14.

Classification of chemo notrec nodes: 3, 7, 11, 15.

Variables in favor of chemo: X1564656, X1558685, X203865, X205204

Variables not in favor of chemo: X1558745, X206645, X1561262

The diabetic retinopathy data set contains 197 patients with 61% censoring. According to the paper by Blair et al., the experiment only those patients who have experienced diabetes for more than 10 years are included in the study. Visual functions were recorded and opthalmological examination was performed at each visit (Blair et al., 1980). Each patient had one eye randomized to laser treatment and the other eye received no treatment. Therefore, each patient has two observations in the data set. The event of interest in this study is the time from initial treatment to the time when visual acuity dropped below 5/200 two visits in a row. Survival times in this dataset are the actual time to vision loss in month, minus the minimum possible time to event (6.5 months). The following variables are included in our analysis: survival time, status (loss of vision/censored), treatment (control/treatment), age type (juvenile/adult), laser type (argon/xenon), risk score for the eye (6 or higher indicates high-risk).
The treatment effect was statistically significant using the Cox model [95% CI (0.33,0.64), $p=2\cdot 10^{-6}$]. With all the data from Series 37745, Figure 6 illustrates the unpruned rpart tree. When we compare the prediction for the whole data set with the actual treatment, the prediction accuracy is 0.61.

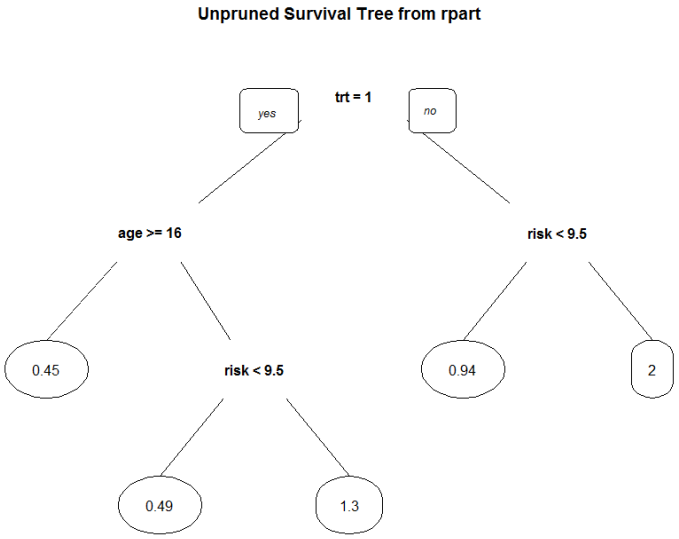

Figure 6: Unpruned Survival Tree for Diabetic Retinopathy Data set

Table 4: Illustration of the Split for the Unpruned Tree using Retinopathy data set

Node Number	Number of laser treatment	Number of non-laser treatment	Total	Prediction
4	94	0	94	Laser recommended
6	0	96	96	Laser not recommended
7	0	101	101	Laser recommended
8	50	0	50	Laser recommended

9	53	0	53	Laser not recommended

In the treatment group, people younger than 16 years old have high relative event rates (0.49 in node 8 and 1.3 in node 9) than the older age group (0.45 in node 4). If we divide all subjects into two age groups: young and old, where young is defined as age less than 16, and old is defined otherwise. We drew the Kaplan-Meier curves from the two age groups (Figure 7). The figure shows that young has lower survival probabilities in general, which implies young has higher relative event rate in general. Therefore, we conclude that people younger than 16 years old have higher relative event rate than older people in general.

The high-risk group (risk>9.5) has higher relative event rate in both the treatment group (1.3 in node 9 vs. 0.49 in node 8) and control (2 in node 7 vs. 0.94 in node 6) than the low-risk group. Figure 9 below illustrates the Kaplan-Meier curves for the control and treatment groups. The two curves are clearly separated, with lower survival probability on the control group. Thus, we conclude that the control group has higher relative event rate in general.

Although high-risk group has higher relevant event rate than the low-risk group regardless of treatment (Figure 8), by comparing node 8 with node 6, and node 9 with node 7, the relative event rates are lower in node 8 and 9. This supports our assumption that the laser treatment reduces the relevant event rate in general.

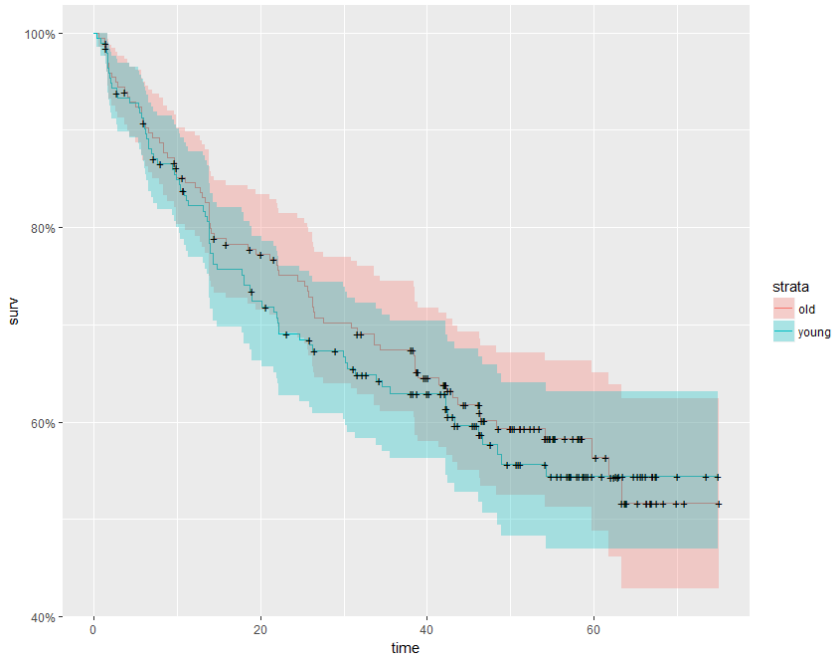

Figure 7: Kaplan-Meier Curves for Old and Young Age Groups

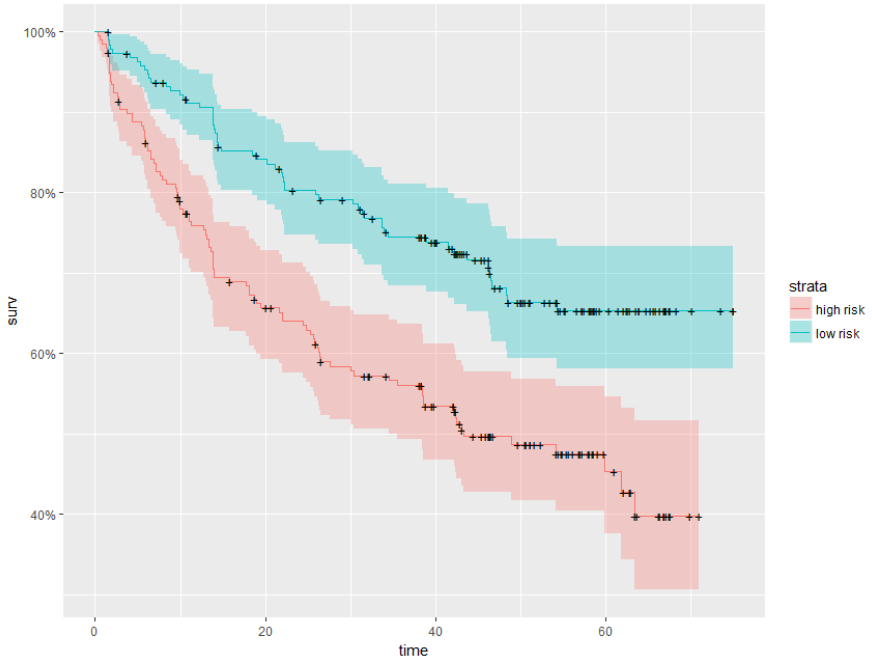

Figure 8: Kaplan-Meier Curves for High and Low Risk Groups

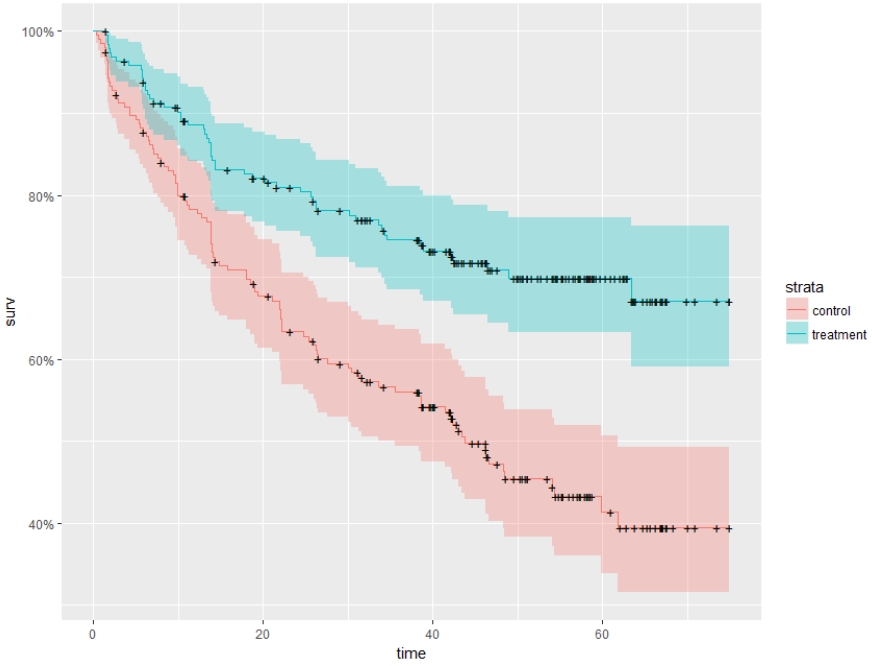

Figure 9: Kaplan-Meier Curves for Control and Treatment Groups

Figure 7 shows the Kaplan-Meier Curves for old and young age groups, old age groups tend to have higher survival time than young age group, and the 95% confidence interval as shown in red and blue colors are overlapping. However, the Kaplan Meier curves for high and low risk group in Figure 8 are well separated, so apparently the survival times in different risk groups are significantly different. Same situation applies

to Figure 9, so we make the hypothesis that treatment effects are different in between the treatment and control groups.

4. Conclusion

From the simulation study, we found that the prediction performance of Design 1 is better than that of Design 2. The reason is because the treatment effect on Design 1 is more significant than that of Design 2. On the other hand, no matter how significant the treatment effect is, we find a way to identify which gene variables are in favor of chemotherapy and which are not. When we applied real data sets to our methods, the 1SE pruning method could result in no split of the tree when the treatment effect is not statistically significant. However, for the Diabetic Retinopathy data set, when the treatment effect is statistically significant using the Cox model [95% CI (0.33,0.64), p=2e-06], the Kaplan-Meier curves for control and treatment groups were well separated although the pruning method will also result in no split of the tree. Patients within different risk groups are also well separated in Kaplan-Meier curves. Based on the simulation results, we conclude that when the treatment effect is more heterogeneous, in other words, more statistically significant, the prediction accuracy tends to be higher. As in the Diabetic Retinopathy data set, it gave a more detailed view of what would be like when the treatment effects are highly heterogeneous. The NCBI data set gave an example of when the treatment effect is not statistically significant.

Acknowledgement

This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the “ICT Consilience Creative Program” (IITP-2015-R0346-15-1007) supervised by the IITP (Institute for Information & communications Technology Promotion).

References

Botling, J., Edlund, K., Lohr, M., Hellwig, B., Holmberg, L., Lambe, M., Berglund, A., Ekman, S., Bergqvist, M., Pontén, F. and König, A., 2013. Biomarker discovery in Non–Small cell lung cancer: Integrating gene expression profiling, meta-analysis, and tissue microarray validation. *Clinical cancer research*, *19*(1). 194-204.

Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A., 1984. *Classification and regression trees*. CRC press.

Blair, A.L., Hadden, D.R., Weaver, J.A., Archer, D.B., Johnston, P.B. and Maguire, C.J., 1980. The 5-year prognosis for vision in diabetes. *The Ulster medical journal*, *49*(2). 139.

Cox, D.R., 1992. Regression models and life-tables. In Breakthroughs in statistics. *Springer New York*. 527-541.

CPMP Working Party on Efficacy on Medicinal Products. Biostatistical methodology in clinical trials in applications for marketing authorizations for medicinal products: Note for guidance. Statistics in Medicine, 14:1659–1682, 1995.

S.-C. Chow and J.-P. Liu. 2004. Design and Analysis of Clinical Trials: Concepts and Methodologies. Hoboken, NJ: Wiley-Interscience.

Su, X., Tsai, C.L., Wang, H., Nickerson, D.M. and Li, B., 2009. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*. 141-158.