Tree Structured Prognostic Model for Hepatocellular Carcinoma

Taerim Lee (1), Hyungjun Cho (2), Hyosuck Lee (3)

(1) Dept of Information Statistics
Korea National Open University
   e-mail : trlee@mail.knou.ac.kr
(2) Dept. of Biostatistics
University of Maryland
   e-mail : hyungjun@stat.wisc.edu
(3) Dept. of Internal Medicine
Seoul National University
   e-mail : hsleemd@snu.ac.kr
1. Background
2. Motivation & Purposes
3. Analytic structure of Tree Structured Model
4. Prognostic Model of Hepatocellular carcinoma
5. Remarks
Background

1. Tree structured regression modeling techniques have been developed using a recursive partitioning algorithm.
2. There has been a strong need on the analyses of the data on the survival rate of patients with HCC who were treated with either with surgical resection or transarterial chemoembolization method.
1. L. Gorden & R. Olshen & (1985) presented tree structured survival analysis in the Cancer Treatment Reports.


4. Lynn A. Sleeper & D.P. Harrington (1990) examined a flexible survival model with *Regression Splines* for covariate effect in Liver Disease.

6. M. LeBlanc & John Crowley (1993) reported survival trees by goodness of split in JASA.


Review of Literature


10. S.H. Um et al. (1997) derived the prognostic index and define the natural stage of HCC using those index.

11. S.H. Um et al. (1998) evaluated the prognosis of HCC in relation the treatment methods.

12. S.K. Han et al. (1998) estimated the survival rate and their affecting factors in patient with HCC.
14. K.M. Kim et al. (2000) studied the survival rate of patient with hepatocellular carcinoma (HCC) who were treated either with surgical resection or transarterial chemoembolization.


1. A variety of powerful modeling technique have been developed for exploring the functional form of effects for HCC.

2. To find the model related to the effects which are changing over time with covariate.

3. To obtain the tree structured prognostic models of HCC patients.
Objectives

1. To identify the effect of prognostic factors of HCC.
2. To quantify the patient characteristics that related to the survival time.
3. To explore the functional form and the relationships of the covariates.
4. To reflect the changing effects over time to the prognostic model.
Data: From 1993 through 1996, 186 patients with HCC in UICC T1-3NOMO and liver function of Child-Pugh class A were enrolled in SNUH.

Variables:

- **Stratification Var:**
  - Surgical Resection group
  - Transarterial chemoembolization group

- **Clinical Var:**
  - sex, age, afp, viral child effect
  - UICC. CLIP # of TACE type, tumor size, lobe, inv.
  - Transformed variables(7)

- **Dependent Var:**
  - prognosis(sur, sc) ⇒ survivalship
# Table 1. Pretreatment Base-Line Characteristics of the Patients

<table>
<thead>
<tr>
<th>CHARACTERISTIC</th>
<th>SURGICAL TETSECTION (N=91)</th>
<th>TRANSCATHETER ARTERIAL CHEMOEMBOLIZATION (N=91)</th>
<th>P VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age-yr</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean ± SD</td>
<td>50 ± 10</td>
<td>56 ± 10</td>
<td>0.0002+</td>
</tr>
<tr>
<td>Male sex – no. (%)</td>
<td>76(84)</td>
<td>79(87)</td>
<td>0.532+</td>
</tr>
<tr>
<td>Serum alpha-fetoprotein – no. (%)</td>
<td></td>
<td></td>
<td>0.756+</td>
</tr>
<tr>
<td>&lt; 400 ng/ml</td>
<td>58(64)</td>
<td>60(66)</td>
<td></td>
</tr>
<tr>
<td>≥ 400 ng/ml</td>
<td>33(36)</td>
<td>31(34)</td>
<td></td>
</tr>
<tr>
<td>Viral marker-no. (%)</td>
<td></td>
<td></td>
<td>0.046</td>
</tr>
<tr>
<td>Hepatitis B virus</td>
<td>71(78)</td>
<td>61(67)</td>
<td></td>
</tr>
<tr>
<td>Hepatitis C virus</td>
<td>4(4)</td>
<td>16(18)</td>
<td></td>
</tr>
<tr>
<td>Hepatitis B and C virus</td>
<td>2(2)</td>
<td>1(1)</td>
<td></td>
</tr>
<tr>
<td>NBNC&amp;</td>
<td>13(14)</td>
<td>10(11)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>1(1)</td>
<td>3(3)</td>
<td></td>
</tr>
</tbody>
</table>
### Table 1. Pretreatment Base-Line Characteristics of the Patients

<table>
<thead>
<tr>
<th>CHARACTERISTIC</th>
<th>SURGICAL SECTION (N=91)</th>
<th>TRANSCATHETER ARTERIAL CHEMOEMBOLIZATION (N=91)</th>
<th>P VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Okuda stage – no.(%)</td>
<td></td>
<td></td>
<td>0.155+</td>
</tr>
<tr>
<td>Okuda</td>
<td>91(100)</td>
<td>89(98)</td>
<td></td>
</tr>
<tr>
<td>Okuda</td>
<td>0(0)</td>
<td>2(2)</td>
<td></td>
</tr>
<tr>
<td>UICC T stage-no.(%)</td>
<td></td>
<td></td>
<td>0.003</td>
</tr>
<tr>
<td>UICC T1</td>
<td>17(19)</td>
<td>12(13)</td>
<td></td>
</tr>
<tr>
<td>UICC T2</td>
<td>63(69)</td>
<td>49(54)</td>
<td></td>
</tr>
<tr>
<td>UICC T3</td>
<td>11(12)</td>
<td>30(33)</td>
<td></td>
</tr>
<tr>
<td>CLIP scoring-no.(%)</td>
<td></td>
<td></td>
<td>0.125</td>
</tr>
<tr>
<td>CLIP1</td>
<td>54(59)</td>
<td>39(43)</td>
<td></td>
</tr>
<tr>
<td>CLIP2</td>
<td>27(30)</td>
<td>34(37)</td>
<td></td>
</tr>
<tr>
<td>CLIP3</td>
<td>8(9)</td>
<td>16(18)</td>
<td></td>
</tr>
<tr>
<td>CLIP4</td>
<td>2(2)</td>
<td>1(1)</td>
<td></td>
</tr>
<tr>
<td>CLIP5</td>
<td>0(0)</td>
<td>1(1)</td>
<td></td>
</tr>
<tr>
<td>Lipiodol retention pattern-no.(%)</td>
<td></td>
<td></td>
<td>0.063</td>
</tr>
<tr>
<td>Compact</td>
<td>56(62)</td>
<td>56(62)</td>
<td></td>
</tr>
<tr>
<td>Non-compact</td>
<td>30(33)</td>
<td>35(38)</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>5(5)</td>
<td>0(0)</td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Changes in alpha-fetoprotein after Initial Treatment in alpha-fetoprotein Secreting HCC

<table>
<thead>
<tr>
<th>CHARACTERISTIC</th>
<th>SURGICAL RESSECTION</th>
<th>TRANSCATHETER ARTERIAL CHEMOEMBOLIZATION</th>
<th>P VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serum alpha-fetoprotein in total alpha-fetoprotein secretion HCC patients</td>
<td>N=73</td>
<td>N=64</td>
<td>0.001+</td>
</tr>
<tr>
<td>No.(%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decreased&gt;50%</td>
<td>61(84)</td>
<td>33(52)</td>
<td></td>
</tr>
<tr>
<td>Decreased 25-50%</td>
<td>7(10)</td>
<td>10(16)</td>
<td></td>
</tr>
<tr>
<td>Stable</td>
<td>3(4)</td>
<td>13(20)</td>
<td></td>
</tr>
<tr>
<td>Increased≥25%</td>
<td>2(3)</td>
<td>8(13)</td>
<td></td>
</tr>
<tr>
<td>Serum alpha-fetoprotein in alpha-fetoprotein secreting HCC patients with compact Lipiodol retention</td>
<td>N=41</td>
<td>N=40</td>
<td>0.137+</td>
</tr>
<tr>
<td>No.(%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decreased&gt;50%</td>
<td>34(83)</td>
<td>25(63)</td>
<td></td>
</tr>
<tr>
<td>Decreased 25-50%</td>
<td>4(10)</td>
<td>6(15)</td>
<td></td>
</tr>
<tr>
<td>Stable</td>
<td>1(2)</td>
<td>6(15)</td>
<td></td>
</tr>
<tr>
<td>Increased≥25%</td>
<td>2(5)</td>
<td>3(8)</td>
<td></td>
</tr>
</tbody>
</table>
Analysis

Survival Analysis - Cox regression

Tree Structured Prognostic Model

Comparison between survival groups
Proportional Hazards Model (Cox 1972)

\[ Z_i : \text{Survival time}, \quad C_i : \text{Censoring time} \]
\[ Y_i = \min(Z_i, C_i), \quad \delta_i = I(Z_i \leq C_i) \]
\[ X_i = (X_{i1}, \ldots, X_{ix})' : \text{Vector of Covariate} \]

If \( \check{\gamma}(y|X) \) is the hazard rate at time \( y \) for an individual with Risk factor \( X \), the Cox proportional hazards Model is
\[
\check{\gamma}(y|X) = \check{\gamma}_0(y)\exp\left( \sum_{k=1}^{p} \hat{\gamma}_k X_k \right)
\]
Fig. 1 Survival Curves for HCC according to Treatment Group

Variables in the Model Exp(B)
CHILD 2.67
CLID 1.43
TX 0.64
EFFECT 0.54

KNOUIMA 2003
Fig. 2 Survival Curves for HCC according to Tumor Stage
Fig. 3 Survival Curves for HCC according to CLIP Stage
Fig. 4 Survival Curves for HCC according to Lipidol Pattern
Fig. 5 Survival Curves for HCC according to Child Score
tree structured prognostic model with effective covariate

CART uses a decision tree to display how data may be classified or predicted.

automatically searches for important relationships and uncovers hidden structure even in highly complex data.
Fig. 6 Tree Structured Model based on HCC data

CART

186
147(0)
39(1)

39
11(0)
28(1)

AFP ≤ 8.0

124
76(0)
48(1)

CHILD ≤ 5.5

23
23(0)
0(1)

29
24(0)
5(1)

95
52(0)
43(1)

TAENUM ≤ 7.5

84
50(0)
34(1)

TAENUM ≤ 1.5

38
15(0)
23(1)

IMA 2003

95
52(0)
43(1)

AGE ≤ 713

11
2(0)
9(1)

29
24(0)
5(1)

1

11
2(0)
9(1)

1

Sensitivity 90.0%
Specificity 30.3%
Total 65.6%

1. AFP 100.0
2. TAENUM 88.8
3. CHILD 61.2
4. AGE 42.2
5. TX 32.1

KNOU
Fig. 7  Tree Structured Model for TACE group of HCC data

CART

Sensitivity 71.7%
Specificity 85.4%
Total 78.7%

1. TAENUM 100.0
2. AFP 87.7
3. CHILD 72.3
4. SIZE 59.4
5. INV 59.0
6. CLIP 45.5
Fig. 8 Tree Structured Model based for SC group of HCC (Surgical Resection Group)

CART

- **AFP ≤ 8.0**
  - 92
  - 73(0)
  - 19(1)

- **TAENUM ≤ 7.5**
  - 73
  - 58(0)
  - 15(1)

- **AGE ≤ 639**
  - 1
  - 7(0)
  - 6(1)

- **AGE > 639**
  - 0
  - 9
  - 6(0)
  - 3(1)

**Statistics**

- **Sensitivity**: 89.1%
- **Specificity**: 67.9%
- **Total**: 82.6%

**Variables**

1. AFP 100.0
2. TAENUM 72.6
3. CHILD 31.1
4. AGE 7.7
5. TX 5.3

**Notes**

IMA 2003
Tree structured prognostic model with effective covariate:

FACT employs statistical hypothesis test to select a variable for splitting each node and then uses discriminant analysis to find the split point. The size of the tree is determined by a set of stopping rules.
QUEST is a new classification tree algorithm derived from the FACT method. It can be used with univariate splits or linear combination splits. Unlike FACT, QUEST uses cross-validation pruning. It distinguishes from other decision tree classifiers is that when used with univariate splits the classifier performs approximately unbiased variable selection.
Survival Tree with Unbiased Detection of Interaction

STUDI is a tree-structured regression modeling tool. It is easy to interpret predict survival value for new case. Missing values can easily be handled and time dependent covariates can be incorporated.
Let the survival function for a covariate $X_i$ be

$$S(t|\psi, x = x_i) = \exp\left(-\Lambda(t, \psi) \exp\beta' x_i\right)$$

Where $\Lambda(t, \psi)$ is the cumulative baseline hazard rate. Then median survival time for and individual $i$ is defined as $\tilde{y}_i = \inf y|\psi, S(t|\psi) \leq 0.5$.

and the cost at a node $t$ be is defined as

$$R(t) = \sum_{i=1}^{n} |y_i - \tilde{y}_i|$$
Modified Cox-Snell (MCS) residuals: for $i=1,...,n$

$$\hat{\Lambda}_0(Y_i) \exp \{\hat{\beta}'X_i\} + 0.693(1 - \delta_i)$$

Where $\hat{\Lambda}_0$ is the estimator of the cumulative baseline hazard function.
Covariate Type

1. n-covariate: Ordered numerical predictor used for fitting and splitting.

2. f-covariate: Ordered numerical predictor used for fitting only.

3. s-covariate: Ordered predictor used for splitting only.

4. c-covariate: Categorical predictor for splitting only
1. Fit a model to \( n \) and \( f \) covariates in the node.

2. Obtain the modified Cox-Snell residuals.

3. Perform a curvature test for each of \( n \)-s-and \( c \)-covariates.

4. Perform a interaction test for each pair of \( n \)-s-and \( c \)-covariates.

5. Select the covariate which has the smallest \( p \)-value.
Bootstrap Calibration

• Bootstrap estimation of a bias correction coefficient
  - Let \((Y^*, \delta^*)\) be a bootstrap sample drawn from the elements in \((Y, \delta)\).
  - Fit a model to \((Y^*, \delta^*, X1, \ldots, Xp)\), using only the n- and f-covariates.
  - Compute the p-values from the \(\chi^2\) curvature and interaction tests.
  - Convert each p-values into an absolute normal variate \(z^*\).
  - For \(r>1\), if \(r \max \{z_n^*, z_{nn}^*\} \geq \max\{z_s^*, z_c^*, z_{ss}^*, z_{cc}^*, z_{ns}^*, z_{nc}^*, z_{sc}^*\}\), an n-covariate is selected.
  - Repeat the above steps over a grid of \(r\) values to estimate \(\pi(r)\), the probability that an n-covariate is selected.
  - Find \(r_0\) such that \(\pi(r_0)\) equals the proportion of n-covariates of the total number of covariates used for spitting in the sample. Interpolate linearly if necessary.
How to select the covariate with the smallest p-value?

1. If the smallest p-value is from a curvature test
   - select the covariate which has the smallest p-value

2. If the smallest p-value is from an interaction test
   - for a pair of n-covariates, select the one which yields the smaller total costs of two subnodes its parent node is split by the sample median.
   - for s or c covariates select the one with the smaller curvature p-value.
   - between n-covariates and s or c covariate select the s or c covariate.
1. Fit a model to \((Y, \delta, X_1 \ldots X_p)\) using only the n and f covariates.

2. Let \(Z' = \max\{r^0Z^n, r^0Z^{nn}, Z^s, Z^c, Z^{ss}, Z^{cc}, Z^{ns}, Z^{nc}, Z^{sc}\}\) select the split covariate using previous selection algorithm.
Split Points or Split Sets

For an ordered covariate, use exhaustive search, it use sample median or sample quantiles.

For a categorical covariate, use a shortcut algorithm given in Breiman et al.

Pruning

1. Cost complexity pruning technique of CART.

2. Choice of tree pruning by a test sample or by cross-validation.
Fig. 9 Tree-Structured Survival Model for HCC

- **Treatment = Surgery**
  - **pafp ≤ 3810**
    - **child ≤ 7**
      - **182**
      - **3**
        - **tareff = non-compact**
          - **12**
            - **23**
              - **y = 27.7**
          - **9**
            - **y = 50.4**
      - **5**
        - **y = 7.3**
  - **147**
    - **pafp ≤ 78**
      - **80**
        - **17**
          - **15**
            - **y = 28.8**
        - **16**
          - **65**
            - **32**
              - **49**
                - **y = 53.5**
            - **130**
              - **65**
                - **34**
                  - **y = 80.2**
        - **17**
          - **15**
            - **y = 73.0**
    - **8**
      - **80**
        - **16**
          - **33**
            - **16**
              - **y = 73.0**
          - **130**
            - **65**
              - **34**
                - **y = 80.2**
        - **17**
          - **15**
            - **y = 50.6**
    - **524**
      - **9**
        - **y = 65.1**
      - **8**
        - **y = 82.3**
    - **525**
      - **8**
        - **y = 11.9**
      - **67**
        - **43**
          - **y = 50.4**
        - **13**
          - **y = 43.5**
  - **taeff = non-compact**
    - **y = 19.1**
    - **36**
      - **24**
        - **19**
          - **y = 43.5**
        - **11**
          - **y = 19.1**
    - **67**
      - **13**
        - **y = 65.1**
      - **8**
        - **y = 82.3**

IMA 2003
Fig. 10 Scatterplot or Boxplots of the MCS Residues
Fig. 11 Survival Curves for HCC according to Node Groups
Fig. 12  Tree-Structured Survival Model for HCC with Surgery

1. If Pafp $\leq 5.0$
2. If $\tilde{y} = 75.2$
3. If $\tilde{y} = 38.4$
Fig. 13 Survival Curves for HCC with Surgery according to pafp Score
Fig. 14 Tree-Structured Survival Model for HCC with TACE

- **Node 1:**
  - Condition: $\text{taen} \leq 7$
  - Outcome: $91$

- **Node 2:**
  - Condition: $\text{pafp} \leq 36.0$
  - Branch: $65$
    - Sub-node 4: $\tilde{y} = 36.1$
    - Sub-node 5: $\tilde{y} = 20.0$

- **Node 3:**
  - Outcome: $26$
  - Sub-node: $\tilde{y} = 57.9$
Fig. 15 Survival Curves for HCC with TACE according to Node Groups
1. The effects of prognostic factors of HCC were identified with the tree structured survival model.

2. The factors CHILD, the number of TACE treatment, AFP, PAFP level were common to all the models, the post AFP was specifically important to the survival time of surgical resection group and TACE treatment group.

3. Tree structured survival model with minimum SAD provide the critical values of splitting covariates for each node.