PBC 자료

- Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984.
- 주요변수 : 생존시간, trt(D-penicillamine and placebo), age, sex 외에 임상, 생화학 지표

```
filename inf 'c:\kim\yes\myweb\survival\pbc.dat';
data pbc ;
    infile inf obs=312 ;
    input n time delta trt age sex ascites hepa spider edema
bili chol
      albumin copper phosphat sgot trig platelet proth stage;
run;
data a;
   set pbc;
   * create dummy variables for edema levels ;
   edema0_5=(edema=0.5);
   edema1=(edema=1) ;
   * change the trt value 1,2 to 0,1 ;
   trt=-trt+2 ;
run;
proc phreg data=A;
   model time*delta(0)=trt ;
run;
```

Model Information

Data Set	WORK.A
Dependent Variable	time
Censoring Variable	delta
Censoring Value(s)	0
Ties Handling	BRESLOW

Number of Observations Read	312
Number of Observations Used	312

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	1279.960	1279.858
AIC	1279.960	1281.858
SBC	1279.960	1284.686

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
312	125	187	59.94

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Testing Global Null Hypothesis: BETA=0 → 모든 BETA=0 ?

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	0.1017	1	0.7498
Score	0.1017	1	0.7498
Wald	0.1015	1	0.7500

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
trt	1	0.05709	0.17916	0.1015	0.7500	1.059

Trt=1(D-penicillamine)이 0에 비해서 위험율이 높지만 통계적인 유의성은 없다

```
proc phreg data=A;
      model time*delta(0)=sex age edema trt bili  ;
run;
```

Sex=0 for male
1 for Female

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
sex	1	-0.53783	0.24510	4.8152	0.0282	0.584
age	1	0.03569	0.00919	15.1007	0.0001	1.036
edema	1	1.64180	0.29641	30.6804	<.0001	5.164
trt	1	0.04519	0.18840	0.0575	0.8104	1.046
bili	1	0.12862	0.01451	78.5316	<.0001	1.137

남성이, 연령이 많을수록 edema 점수가 높을수록 trt=1 이 bili level 높을수록 => 위험율이 커진다.

```
proc phreg data=B;
  model time*delta(0)=sex age edema
        trt  stage2 stage3 stage4 logalb logproth
logbili ;
run;
```

남성이, 연령이 많을수록, edema 많을수록 trt=0이, stage가 후기일수록 albumin level 높을수록 , prothrombin time이 길수록, bili level이 높을 수록 위험이 커진다.

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
sex	1	-0.42994	0.25705	2.7975	0.0944	0.651
age	1	0.02671	0.00950	7.9040	0.0049	1.027
edema	1	0.82139	0.30140	7.4272	0.0064	2.274
trt	1	-0.09556	0.18591	0.2642	0.6073	0.909
stage2	1	1.49319	1.05194	2.0149	0.1558	4.451
stage3	1	1.64658	1.02732	2.5689	0.1090	5.189
stage4	1	1.83441	1.02853	3.1810	0.0745	6.261
logalb	1	-2.78855	0.76963	13.1280	0.0003	0.062
logproth	1	3.12924	1.21316	6.6534	0.0099	22.857
logbili	1	0.83873	0.10175	67.9470	<.0001	2.313

```
proc phreg data=B;
  model time*delta(0)=sex age edema
        trt  stage2 stage3 stage4 logalb logproth
logbili /selection=stepwis;
run;
```

The PHREG Procedure

Model Information

Data Set	WORK.B
Dependent Variable	time
Censoring Variable	delta
Censoring Value(s)	0
Ties Handling	BRESLOW

Number of Observations Read	312
Number of Observations Used	312

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
312	125	187	59.94

Step 1. Variable logbili is entered. The model contains the following explanatory variables: logbili

.....

Step 2. Variable logalb is entered. The model contains the following explanatory variables: logalb logbili

.....

Step 3. Variable age is entered. The model contains the following explanatory variables: age logalb logbili

.....

Step 4. Variable logproth is entered. The model contains the following explanatory variables: age logalb logproth logbili

.....

Step 5. Variable edema is entered. The model contains the following explanatory variables: age edema logalb logproth logbili

NOTE: No (additional) variables met the 0.05 level for entry into the model.

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
age	1	0.03327	0.00866	14.7565	0.0001	1.034
edema	1	0.78470	0.29913	6.8814	0.0087	2.192
logalb	1	-3.05332	0.72408	17.7818	<.0001	0.047
logproth	1	3.01567	1.02380	8.6764	0.0032	20.403
logbili	1	0.87921	0.09873	79.2980	<.0001	2.409

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	logbili		1	159.0619	.	<.0001
2	logalb		2	34.1169	.	<.0001
3	age		3	18.2606	.	<.0001
4	logproth		4	13.1135	.	0.0003
5	edema		5	7.0129	.	0.0081

$$\log \hat{h}(t) = \log h_0(t) + 0.03327 \bullet AGE + 0.78470 \bullet EDEMA - 3.05332 * \log alb$$
$$+ 3.01567 \bullet \log proth + 0.87921 \bullet \log alb$$

문) 다른 조건들이 동일하다고 할 때 연령이 10세 증가하면 위험율이 얼마나 증가하는가 ?
답) 로그 위험율이 0.03327*10 증가한다.
위험율은 exp(0.03327*10)=1.395배, 즉 39.5% 증가한다.

문) 다른 조건들이 동일하다고 할 때 연령이 20세 증가하면 위험율이 얼마나 증가하는가 ?
답) 로그 위험율이 0.03327*20 증가한다.
위험율은 exp(0.03327*20)=1.945배, 즉 94.5% 증가한다.

문) 다른 조건들이 동일하다고 할 때 Edema=1 인 경우는 0인 경우에 비해서 위험율이 얼마나 증가하는가 ?
답) 로그위험율이 0.7847 증가한다.
위험율이 exp(0.7847)=2.19 배 증가한다.

문) 다른 조건들이 동일하다고 할 때 Edema=1 인 경우는 0.5 인 경우에 비해서 위험율이 얼마나 증가하는가 ?
답) 로그위험율이 0.7847*0.5 증가한다.
위험율이 exp(0.7847)=1.48 배 증가한다.

$$\log \hat{h}(t) = \log h_0(t) + 0.03327 \bullet AGE + 0.78470 \bullet EDEMA - 3.05332 * \log alb$$
$$+ 3.01567 \bullet \log proth + 0.87921 \bullet \log alb$$

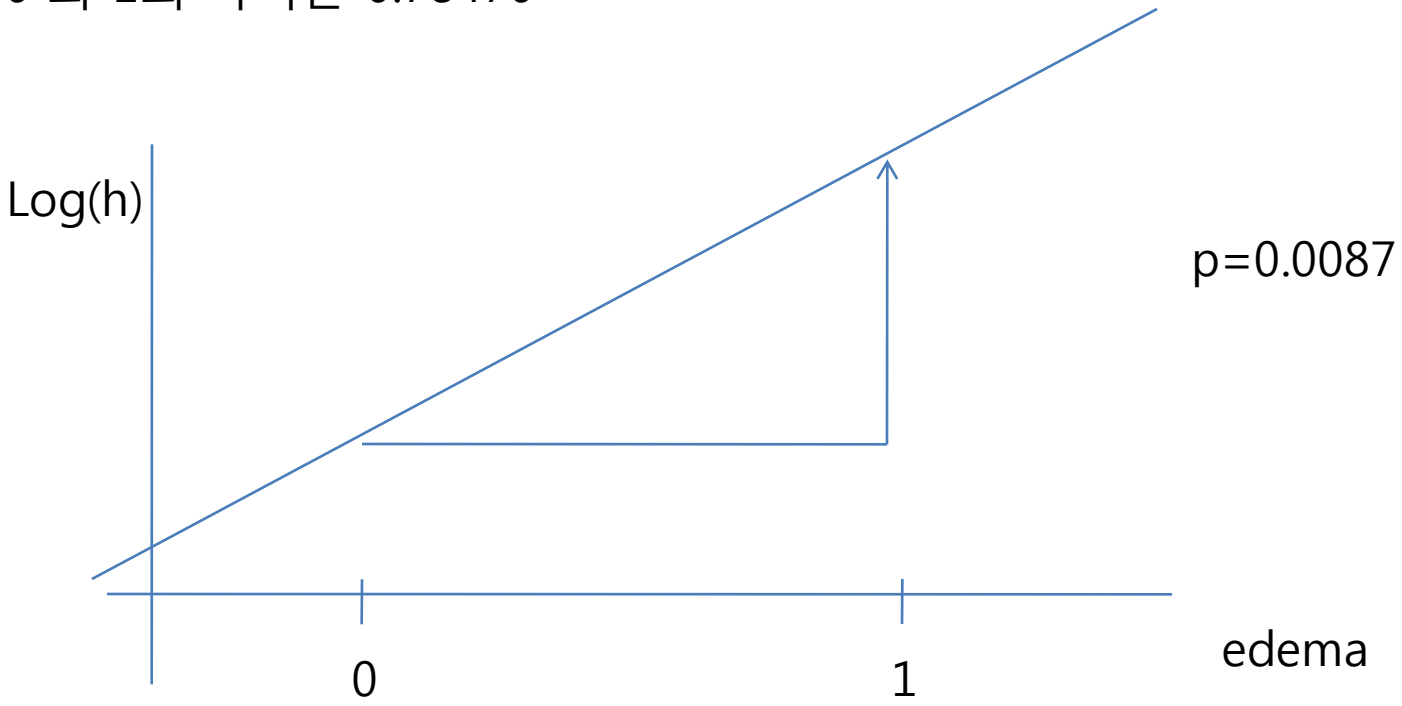
문) 다른 조건들이 동일하다고 할 때 Edema=0.5 인 경우는 0 인 경우에 비해서 위험율이 얼마나 증가하는가 ?
답) 로그위험율이 0.7847*0.5 증가한다.
위험율이 exp(0.7847)=1.48 배 증가한다.

```
proc phreg data=B;
  model time*delta(0)= age edema0_5 edema1
        logalb logproth logbili ;
run;
```

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
age	1	0.03404	0.00868	15.3719	<.0001	1.035
edema0_5	1	0.12112	0.27498	0.1940	0.6596	1.129
edema1	1	0.91055	0.30718	8.7867	0.0030	2.486
logalb	1	-3.07039	0.72567	17.9023	<.0001	0.046
logproth	1	3.00085	1.03325	8.4349	0.0037	20.103
logbili	1	0.87707	0.09920	78.1772	<.0001	2.404

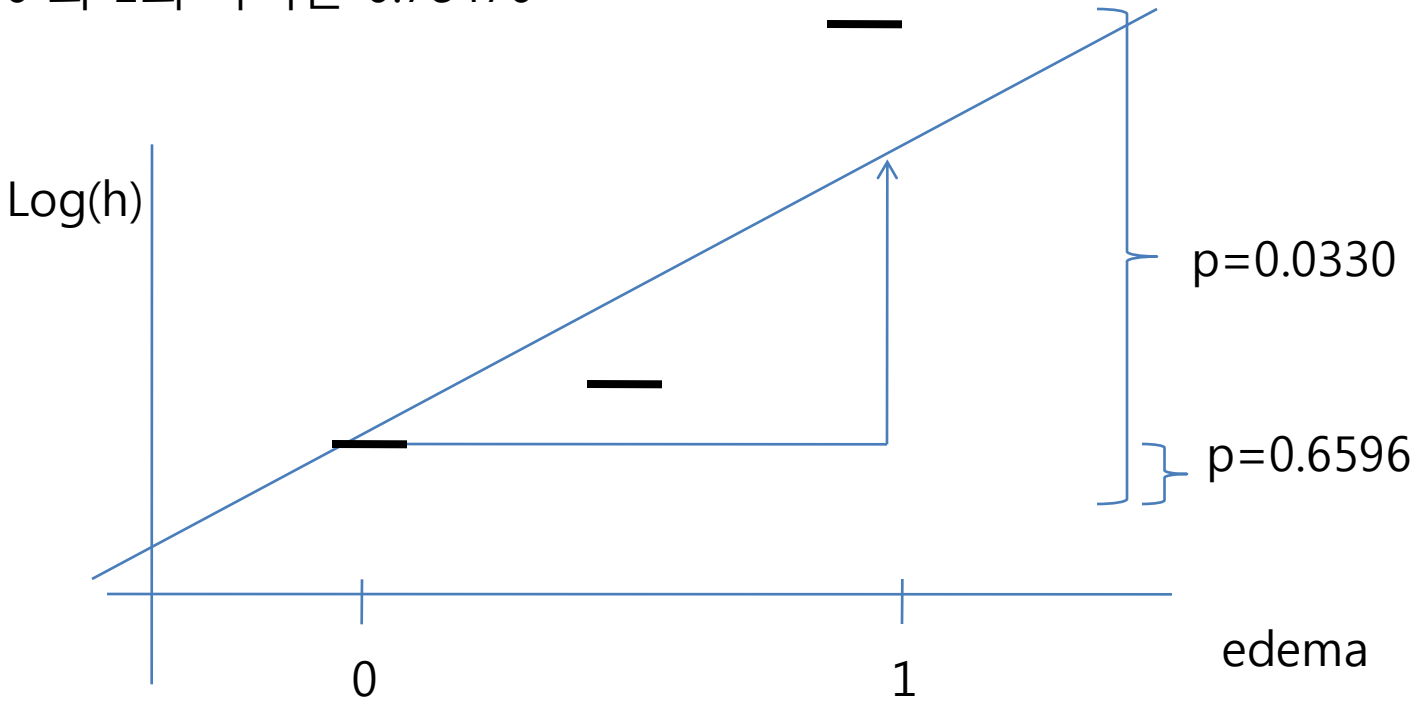
이 모형에서
edema 0.5와 0(reference)의 차이는 0.12112
edema 1과 0(reference)의 차이는 0.91055 그러므로
edema 0.5와 1의 차이는 0.91055-0.12112=0.78943

그 전의 모형에서는
edema 0.5와 0의 차이는 0.78470/2=0.39235
edema 0.5와 1의 차이는 0.78470/2=0.39235
Edema 0 과 1의 차이는 0.78470



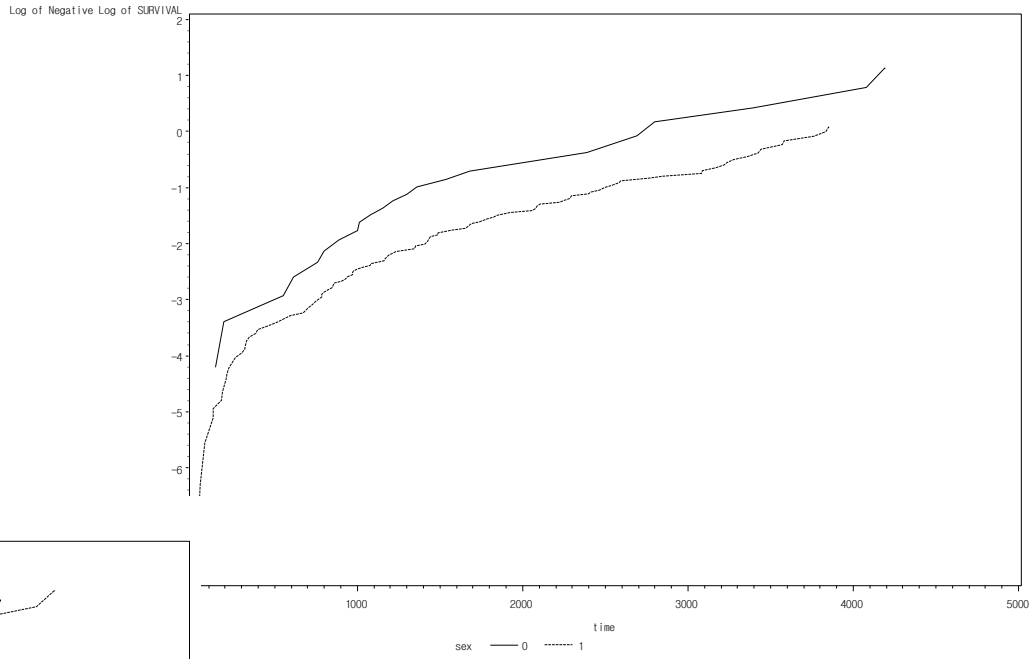
이 모형에서
edema 0.5와 0(reference)의 차이는 0.12112
edema 1과 0(reference)의 차이는 0.91055 그러므로
edema 0.5와 1의 차이는 0.91055-0.12112=0.78943

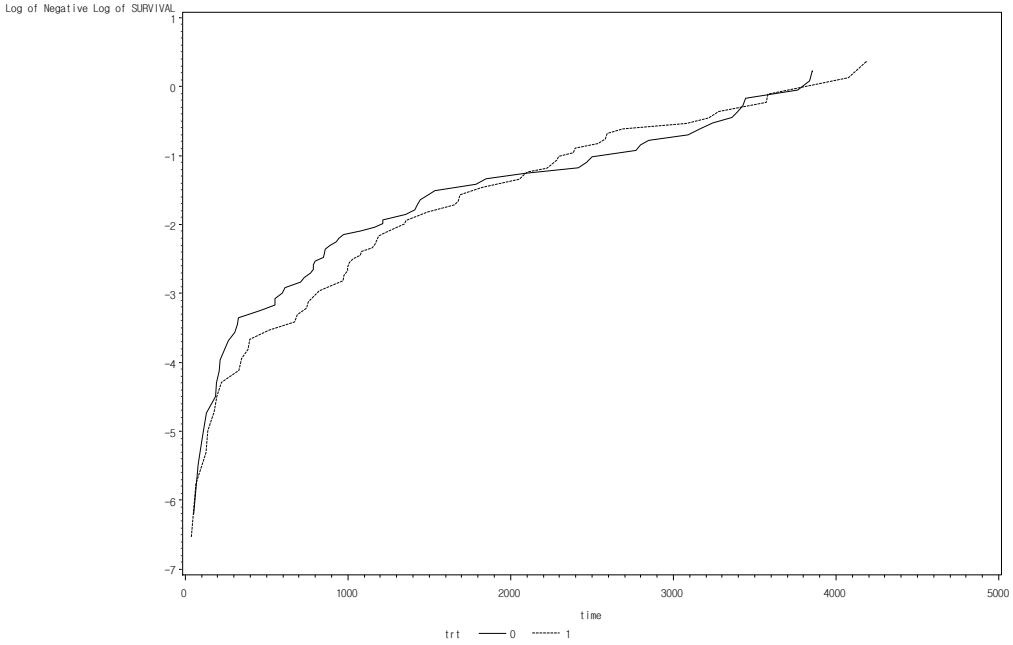
그 전의 모형에서는
edema 0.5와 0의 차이는 0.78470/2=0.39235
edema 0.5와 1의 차이는 0.78470/2=0.39235
Edema 0 과 1의 차이는 0.78470



```
proc phreg data=B;
  model time*delta(0)=trt age edema
        bili  logalb logproth logbili;
  strata sex ;
  baseline out=c survival=s loglogs=lls ;
run;

proc gplot data=c;
    plot lls*time=sex;
      symbol1 i=join c=black line=1 ;
      symbol2 i=join c=black line=2 ;
run;
```

Cox 모형에서의 비례성 검정





The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 JANUARY 4, 2007 VOL. 356 NO. 1

A Five-Gene Signature and Clinical Outcome in Non–Small-Cell Lung Cancer

Hsuan-Yu Chen, M.Sc., Sung-Liang Yu, Ph.D., Chun-Houh Chen, Ph.D., Gee-Chen Chang, M.D., Ph.D., Chih-Yi Chen, M.D., Ang Yuan, M.D., Ph.D., Chiou-Ling Cheng, M.Sc., Chien-Hsun Wang, M.Sc., Harn-Jing Terng, Ph.D., Shu-Fang Kao, M.Sc., Wing-Kai Chan, M.D., Han-Ni Li, M.Sc., Chun-Chi Liu, M.Sc., Sher Singh, Ph.D., Wei J. Chen, M.D., Sc.D., Jeremy J.W. Chen, Ph.D., and Pan-Chyr Yang, M.D., Ph.D.

ABSTRACT

BACKGROUND

Current staging methods are inadequate for predicting the outcome of treatment of non–small-cell lung cancer (NSCLC). We developed a five-gene signature that is closely associated with survival of patients with NSCLC.

METHODS

We used computer-generated random numbers to assign 185 frozen specimens for microarray analysis, real-time reverse-transcriptase polymerase chain reaction (RT-PCR) analysis, or both. We studied gene expression in frozen specimens of lung-cancer tissue from 125 randomly selected patients who had undergone surgical resection of NSCLC and evaluated the association between the level of expression and survival. We used risk scores and decision-tree analysis to develop a gene-expression model for the prediction of the outcome of treatment of NSCLC. For validation, we used randomly assigned specimens from 60 other patients.

From National Taiwan University College of Public Health (H.-Y.C., W.J.C.), National Taiwan University College of Medicine (H.-Y.C., S.-L.Y., C.-L.C., C.-H.W., S.-F.K., H.-N.L., S.S., W.J.C., J.J.W.C., P.-C.Y.), Academia Sinica (C.-H.C, P.-C.Y.), National Taiwan University Hospital (A.Y., W.-K.C., P.-C.Y.), and Advpharma (H.-J.T.) — all in Taipei, Taiwan; and Taichung Veterans General Hospital (G.-C.C., C.-Y.C.) and National Chung-Hsing University (G.-C.C., C.-C.L., J.J.W.C.) — both in Taichung, Taiwan. Address reprint requests to Dr. Yang at the Department of Internal Medicine, National Taiwan University Hospital, No. 7, Chung-Shan S. Rd., Taipei, Taiwan 100, or at pcyang@ha.mc.ntu.edu.tw.

METHODS

We used computer-generated random numbers to assign 185 frozen specimens for microarray analysis, real-time reverse-transcriptase polymerase chain reaction (RT-PCR) analysis, or both. We studied gene expression in frozen specimens of lung-cancer tissue from 125 randomly selected patients who had undergone surgical resection of NSCLC and evaluated the association between the level of expression and survival. We used risk scores and decision-tree analysis to develop a gene-expression model for the prediction of the outcome of treatment of NSCLC. For validation, we used randomly assigned specimens from 60 other patients.

RESULTS

Sixteen genes that correlated with survival among patients with NSCLC were identified by analyzing microarray data and risk scores. We selected five genes (*DUSP6*, *MMD*, *STAT1*, *ERBB3*, and *LCK*) for RT-PCR and decision-tree analysis. The five-gene signature was an independent predictor of relapse-free and overall survival. We validated the model with data from an independent cohort of 60 patients with NSCLC and with a set of published microarray data from 86 patients with NSCLC.

CONCLUSIONS

Our five-gene signature is closely associated with relapse-free and overall survival among patients with NSCLC.

demia Sinica (C.-H.C., P.-C.Y.), National Taiwan University Hospital (A.Y., W.-K.C., P.-C.Y.), and Advpharma (H.-J.T.) — all in Taipei, Taiwan; and Taichung Veterans General Hospital (G.-C.C., C.-Y.C.) and National Chung-Hsing University (G.-C.C., C.-C.L., J.J.W.C.) — both in Taichung, Taiwan. Address reprint requests to Dr. Yang at the Department of Internal Medicine, National Taiwan University Hospital, No. 7, Chung-Shan S. Rd., Taipei, Taiwan 100, or at pcyang@ha.mc.ntu.edu.tw.

Drs. W.J. Chen, J.J.W. Chen, and P.C. Yang contributed equally to this article.

N Engl J Med 2007;356:11-20.

Glossary.

Decision tree: A statistical tool for predicting which patient belongs to which specific class (e.g., good or poor clinical outcome) on the basis of feature information (gene-expression levels), with the use of a recursive-partitioning process and tree-based classification rules.

Gene-expression profiling: Determination of the level of expression of thousands of genes simultaneously by DNA microarray or real-time RT-PCR.

High-risk gene signature: Aberrant expression of a panel of genes in tissue that signifies a high risk of an adverse outcome (relapse or death in patients with cancer).

Independent cohort: An independent group of patients having clinical characteristics similar to those of an original group of patients in a study. The independent cohort is used to confirm the findings of the original study.

Risk gene: A gene for which altered expression in the tissue of interest is associated with an increased risk of an adverse clinical outcome (relapse or death in patients with cancer).

Risk score: A score that predicts the likelihood of an individual patient's survival on the basis of statistical analysis of risk factors (the expression levels of risk genes) associated with survival.

Table 1. Hazard Ratios for Death from Any Cause for the 125 Patients with NSCLC and Results of Validation of the 16-Gene Signature.*

Gene	UniGene Number	Hazard Ratio	P Value†	Correlation Coefficient for Microarray Results vs. Real-Time RT-PCR Results	P Value‡
ERBB3	Hs.118681	1.73	0.03	0.59	<0.001
LCK	Hs.470627	0.43	0.02	0.55	<0.001
DUSP6	Hs.298654	2.12	0.01	0.46	<0.001
STAT1	Hs.470943	0.56	0.02	0.40	<0.001
MMD	Hs.463483	2.50	0.04	0.27	0.006
CPEB4	Hs.127126	1.80	0.02	0.16	0.12
RNF4	Hs.66394	1.91	0.02	0.13	0.18
STAT2	Hs.530595	1.80	0.03	0.15	0.12
NF1	Hs.113577	1.60	0.04	–0.15	0.12
FRAP1	Hs.338207	0.46	0.04	–0.12	0.24
DLG2	Hs.503453	3.75	0.004	–0.09	0.37
IRF4	Hs.401013	1.68	0.03	0.06	0.57
ANXA5	Hs.480653	0.34	0.004	0.06	0.57
HMMR	Hs.72550	1.67	0.04	–0.03	0.79
HGF	Hs.396530	1.66	0.03	0.02	0.82
ZNF264	Hs.515634	1.73	0.01	0.01	0.95

* The hazard ratios are reported for the high-risk signature versus the low-risk signature, as determined by microarray analysis. The first five genes shown were selected for the prediction of survival and used in the decision-tree analysis.
† P values for the hazard ratios were estimated by univariate Cox regression analysis of the microarray data.
‡ P values for the correlation coefficients were estimated by Spearman's rank-correlation test.

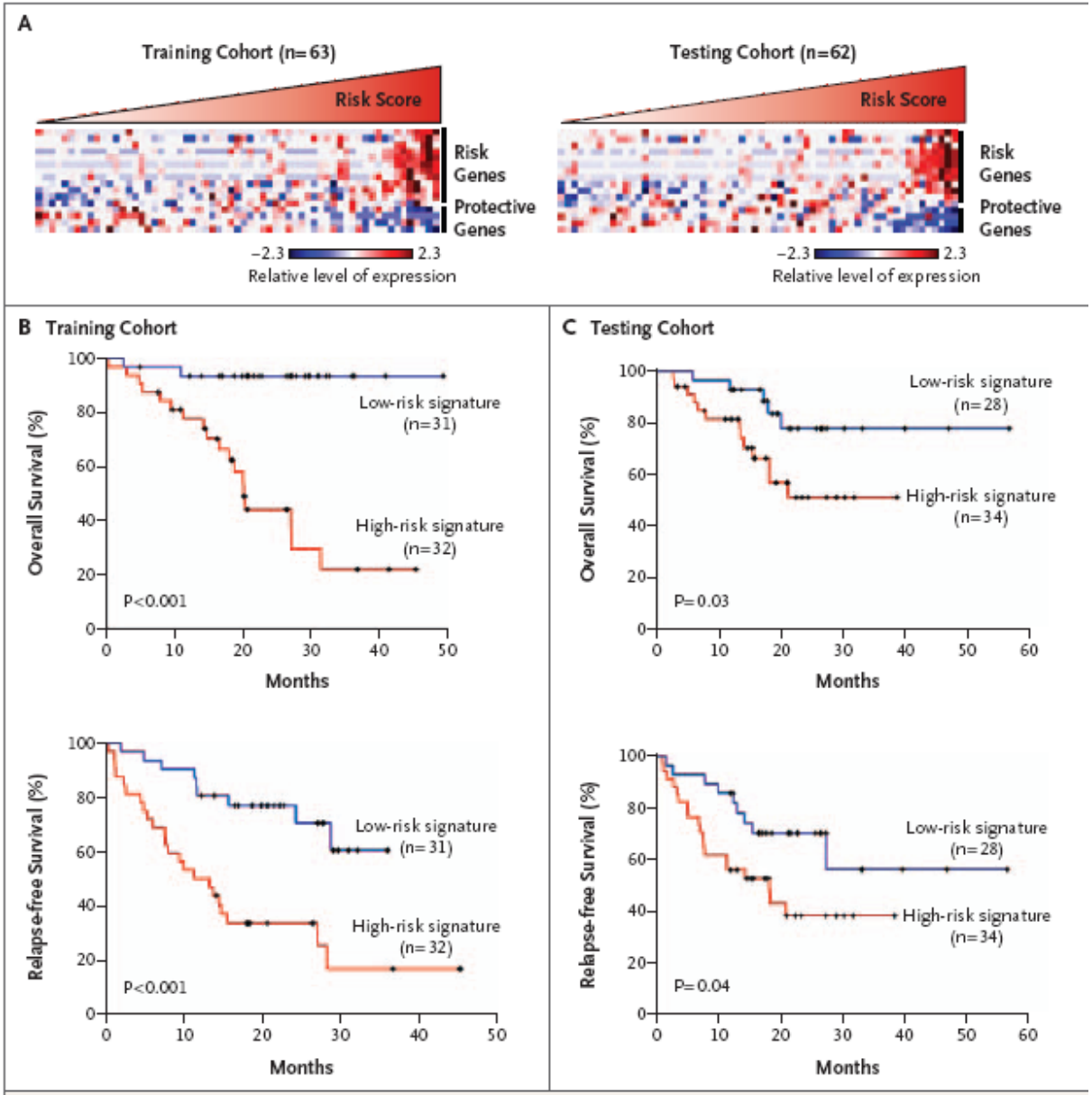


Figure 1. The 16-Gene Signature and Survival of 125 Patients with NSCLC.

Panel A shows the gene-expression profiles of the tumor specimens (according to the color scale shown); each column represents an individual patient. The magnitude of the corresponding risk scores is represented by the slope of the red triangle. Also shown are Kaplan–Meier estimates of overall and relapse-free survival according to the 16-gene microarray signature in the training cohort (Panel B) and the testing cohort (Panel C).

Table 2. Clinical Characteristics of the Original and Validation Cohorts.*

Characteristic	High-Risk Gene Signature	Low-Risk Gene Signature	P Value
Original cohort			
No. of patients	59	42	
Age — yr	65.0±11.6	66.3±10.7	0.54†
Sex — no. of patients (%)			
Male	45 (76)	35 (83)	0.46‡
Female	14 (24)	7 (17)	
Tumor stage — no. of patients (%)			
I or II	29 (49)	30 (71)	0.04‡
III	30 (51)	12 (29)	
Tumor type — no. of patients (%)			
Adenocarcinoma	36 (61)	15 (36)	0.03§
Squamous-cell carcinoma	19 (32)	20 (47)	
Other	4 (7)	7 (17)	
Validation cohort			
No. of patients	34	26	
Age — yr	69.4±9.2	65.3±10.3	0.11†
Sex — no. of patients (%)			
Male	30 (88)	20 (77)	0.31‡
Female	4 (12)	6 (23)	
Tumor stage — no. of patients (%)			
I or II	20 (59)	22 (85)	0.046‡
III	14 (41)	4 (15)	
Tumor type — no. of patients (%)			
Adenocarcinoma	11 (32)	13 (50)	0.19§
Squamous-cell carcinoma	20 (59)	11 (42)	
Other	3 (9)	2 (8)	

* The original cohort consisted of the 101 patients for whom the five-gene signatures from microarray analysis and RT-PCR analysis were significantly correlated. Plus–minus values are means ±SD.
† The P value was calculated by the t-test.
‡ The P value was calculated by Fisher's exact test.
§ The P value was calculated by the chi-square test.

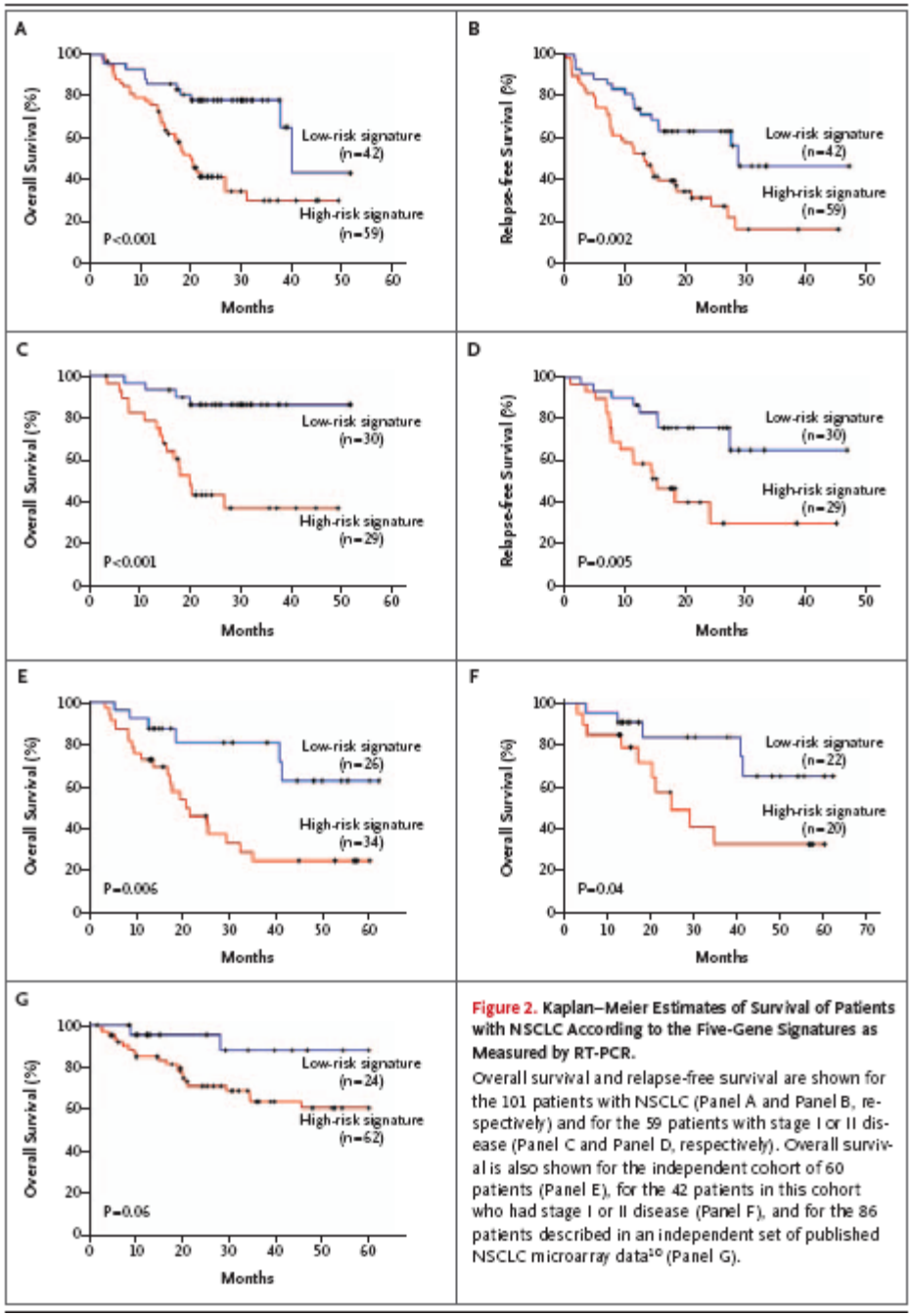


Figure 2. Kaplan–Meier Estimates of Survival of Patients with NSCLC According to the Five-Gene Signatures as Measured by RT-PCR.

Overall survival and relapse-free survival are shown for the 101 patients with NSCLC (Panel A and Panel B, respectively) and for the 59 patients with stage I or II disease (Panel C and Panel D, respectively). Overall survival is also shown for the independent cohort of 60 patients (Panel E), for the 42 patients in this cohort who had stage I or II disease (Panel F), and for the 86 patients described in an independent set of published NSCLC microarray data[10] (Panel G).

Table 3. Hazard Ratios for Death from Any Cause Among Patients with NSCLC, According to Multivariate Cox Regression Analysis.*

Variable	Hazard Ratio (95% CI)	P Value
Original cohort		
High-risk five-gene signature	2.82 (1.38–5.78)	0.005
Tumor stage III	2.13 (1.16–3.93)	0.02
Older age	1.06 (1.03–1.09)	<0.001
Validation cohort		
High-risk five-gene signature	3.36 (1.35–8.35)	0.009
Validation microarray data set		
High-risk five-gene signature	4.36 (1.01–18.76)	0.048
Tumor stage III	7.50 (3.18–17.66)	<0.001

* Variables were selected with a stepwise selection method. The equation used to identify the high-risk five-gene signature is given in the Supplementary Appendix. There were 101 patients in the original cohort (those for whom the five-gene signatures from microarray analysis and RT-PCR analysis were significantly correlated), 60 in the validation cohort, and 86 in the validation data set. CI denotes confidence interval.

SAS PROC PHREG를 이용한 Stepwise Regression

http://support.sas.com/onlinedoc/913/docMainpage.jsp

Data from a study on multiple myeloma in which researchers t reated 65 patients with alkylating agents.

```
data Myeloma;
        input Time VStatus LogBUN HGB Platelet Age LogWBC Frac
              LogPBM Protein SCalc;
        label Time='Survival Time'
              VStatus='0=Alive 1=Dead';
        datalines;
      1.25  1  2.2175   9.4  1  67  3.6628  1  1.9542  12  10
      1.25  1  1.9395  12.0  1  38  3.9868  1  1.9542  20  18
      2.00  1  1.5185   9.8  1  81  3.8751  1  2.0000   2  15
      2.00  1  1.7482  11.3  0  75  3.8062  1  1.2553   0  12
      2.00  1  1.3010   5.1  0  57  3.7243  1  2.0000   3   9
      3.00  1  1.5441   6.7  1  46  4.4757  0  1.9345  12  10
생략
```

```
proc phreg data=Myeloma; model Time*VStatus(0)=LogBUN HG
B Platelet Age LogWBC Frac LogPBM Protein SCalc / selection=s
tepwise slentry=0.25 slstay=0.15 details; run;
```

Model Information

Data Set	WORK.MYELOMA	
Dependent Variable	Time	Survival Time
Censoring Variable	VStatus	0=Alive 1=Dead
Censoring Value(s)	0	
Ties Handling	BRESLOW	

Number of Observations Read	65
Number of Observations Used	65

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
65	48	17	26.15

Analysis of Variables Not in the Model

Variable	Score Chi-Square	Pr > ChiSq
LogBUN	8.5164	0.0035
HGB	5.0664	0.0244
Platelet	3.1816	0.0745
Age	0.0183	0.8924
LogWBC	0.5658	0.4519
Frac	0.9151	0.3388
LogPBM	0.5846	0.4445
Protein	0.1466	0.7018
SCalc	1.1109	0.2919

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
18.4550	9	0.0302

Step 1. Variable LogBUN is entered. The model contains the following explanatory variables:

LogBUN

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	309.716	301.959
AIC	309.716	303.959
SBC	309.716	305.830

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	7.7572	1	0.0053
Score	8.5164	1	0.0035
Wald	8.3392	1	0.0039

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
LogBUN	1	1.74595	0.60460	8.3392	0.0039	5.731

Analysis of Variables Not in the Model

Variable	Score Chi-Square	Pr > ChiSq
HGB	4.3468	0.0371
Platelet	2.0183	0.1554
Age	0.7159	0.3975
LogWBC	0.0704	0.7908
Frac	1.0354	0.3089
LogPBM	1.0334	0.3094
Protein	0.5214	0.4703
SCalc	1.4150	0.2342

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
9.3164	8	0.3163

Step 2. Variable HGB is entered. The model contains the following explanatory variables:

LogBUN HGB

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	309.716	297.767
AIC	309.716	301.767
SBC	309.716	305.509

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	11.9493	2	0.0025
Score	12.7252	2	0.0017
Wald	12.1900	2	0.0023

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
LogBUN	1	1.67440	0.61209	7.4833	0.0062	5.336
HGB	1	-0.11899	0.05751	4.2811	0.0385	0.888

Analysis of Variables Not in the Model

Variable	Score Chi-Square	Pr > ChiSq
Platelet	0.2266	0.6341
Age	1.3508	0.2451
LogWBC	0.3785	0.5384
Frac	1.0491	0.3057
LogPBM	0.6741	0.4116
Protein	0.6592	0.4168
SCalc	1.8225	0.1770

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
5.3635	7	0.6157

Step 3. Variable SCalc is entered. The model contains the following explanatory variables:

LogBUN HGB SCalc

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	309.716	296.078
AIC	309.716	302.078
SBC	309.716	307.692

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	13.6377	3	0.0034
Score	15.3053	3	0.0016
Wald	14.4542	3	0.0023

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
LogBUN	1	1.63593	0.62359	6.8822	0.0087	5.134
HGB	1	-0.12643	0.05868	4.6419	0.0312	0.881
SCalc	1	0.13286	0.09868	1.8127	0.1782	1.142

Step 4. Variable SCalc is removed. The model contains the following explanatory variables:

LogBUN HGB

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	309.716	297.767
AIC	309.716	301.767
SBC	309.716	305.509

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	11.9493	2	0.0025
Score	12.7252	2	0.0017
Wald	12.1900	2	0.0023

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
LogBUN	1	1.67440	0.61209	7.4833	0.0062	5.336
HGB	1	-0.11899	0.05751	4.2811	0.0385	0.888

The PHREG Procedure

Analysis of Variables Not in the Model

Variable	Score Chi-Square	Pr > ChiSq
Platelet	0.2266	0.6341
Age	1.3508	0.2451
LogWBC	0.3785	0.5384
Frac	1.0491	0.3057
LogPBM	0.6741	0.4116
Protein	0.6592	0.4168
SCalc	1.8225	0.1770

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
5.3635	7	0.6157

NOTE: Model building terminates because the variable to be entered is the variable that was removed in the last step.

Summary of Stepwise Selection

Step	Variable Entered	Removed	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	LogBUN		1	8.5164	.	0.0035
2	HGB		2	4.3468	.	0.0371
3	SCalc		3	1.8225	.	0.1770
4		SCalc	2	.	1.8127	0.1782

Cox 모형

- 생존자료에서 다른 변수들의 효과를 보정한 후 trt효과를 볼 수 있는 가장 대표적인 통계모형
- Cox 모형의 결과는 HR (위험율)로 해석함
- 비례가정은 반드시 확인하는 것이 원칙임 -> 비례가정이 만족되지 않을 경우 time dependent covariate approach (예.Extended Cox 모형)

시간형 결과의 비교 (생존분석 이용)
비모수적 방법 (로그순위 검정 이용)

- 단형할당의 경우 총 사건수

$$D = 4\frac{(Z_\alpha + Z_\beta)^2(\theta+1)^2}{(\theta-1)^2}$$

- 예) 검정력 90%를 가지고 위험비 1.75를 95% 유의수준의 양쪽 검정으로 위험비 1로부터 유의하게 감지할 수 있는 검정을 위해서는

$$D=\frac{(1.96+1.282)^2(1.75+1)^2}{(1.75-1)^2}=141\ event \quad <\text{표}7>$$

- 만약 30%가 실험 종료까지 사건이 발생하지 않는다면 (event free) (중도절단비, censoring rate=30%),
  총표본수는 141/0.30=202로 주어진다.

- **표7. 로그 순위 검정을 이용한 비교를 위해 필요한 (두 집단의) 총표본수 표** (위:모수적, 아래:비모수적 방법)

⊿	양측	β=0.1		β=0.2	
		α=0.05	α=0.10	α=0.05	α=0.10
1.25		844 852	688 694	630 636	496 500
1.50		256 262	208 214	192 196	150 154
1.75		134 142	110 116	100 106	80 84
2.00		88 94	72 78	66 70	52 56
2.25		64 72	52 58	48 54	38 42
2.50		50 58	42 48	28 42	30 34

피험자수 계산
시간형 결과의 비교 (생존분석 이용)
모수적 방법 (지수분포 이용)

- 총 사건수 $D = 4\dfrac{(Z_\alpha + Z_\beta)^2}{(\log(\theta))^2}$, $\Delta =$ 위험비

- 예) 검정력 90%를 가지고 위험비 2.0을 95% 유의수준의 양쪽 검정으로 유의하게 감지할 수 있는 검정을 위해서는

$$4\frac{(1.96+1.282)^2}{(\log(2.0))^2} = 87.50 \cong 88 \; event$$가 필요하다.

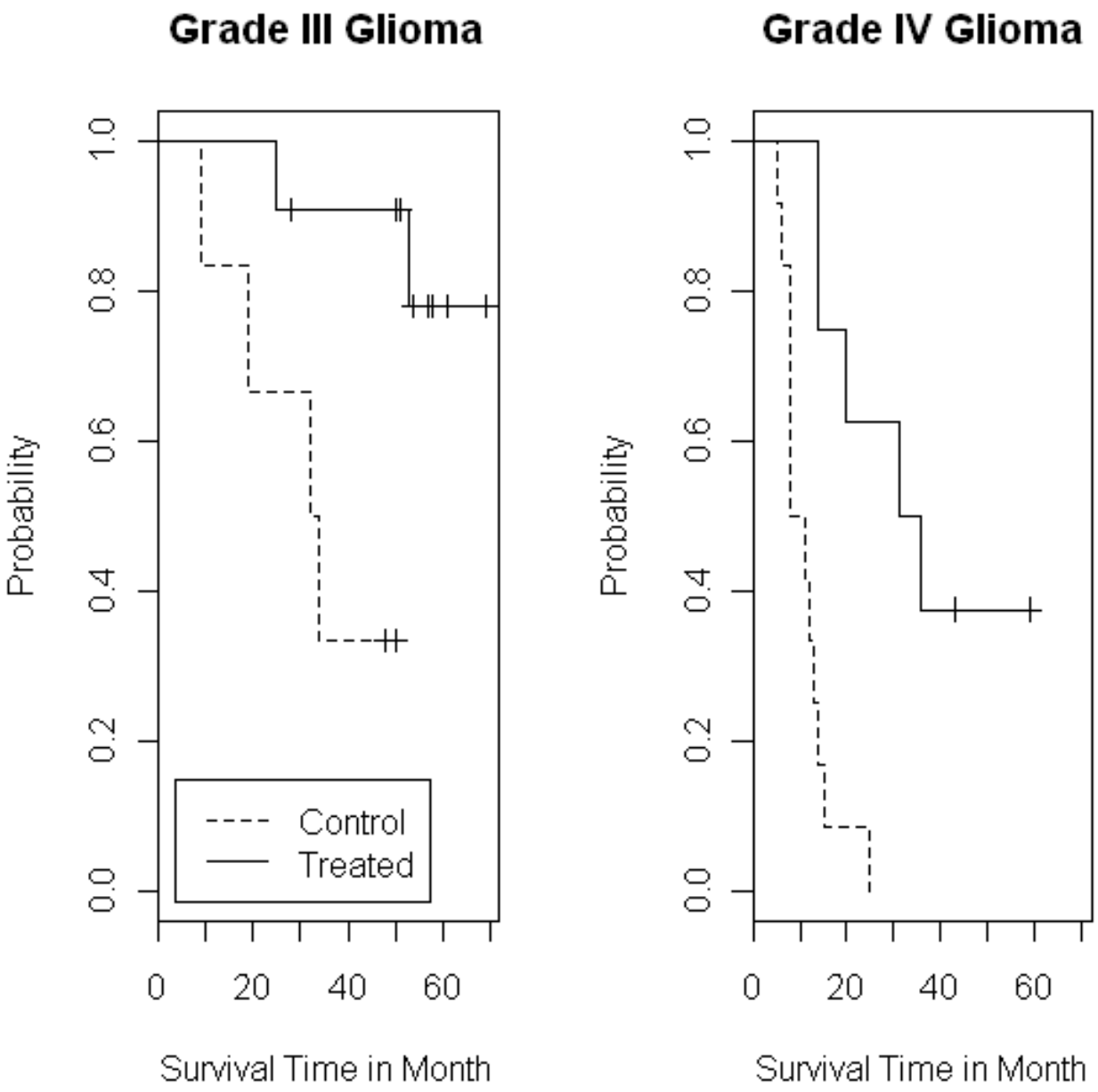
즉 실험군 비교군 각각 44개의사건이 필요하게 된다.

표7 과 동일한 결과

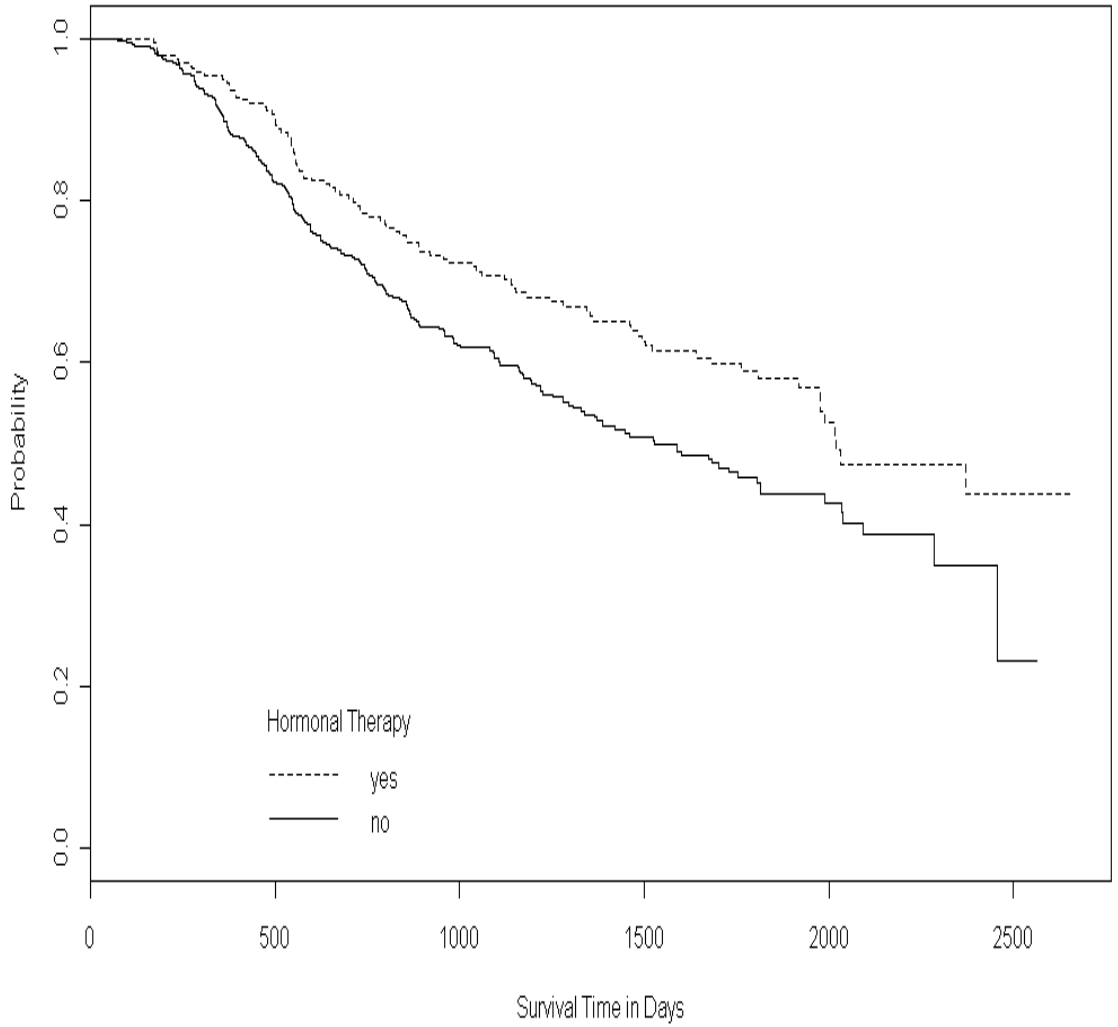
R을 이용한 생존분석

```
> library("HSAUR")
요구된 패키지 lattice를 로드중입니다
요구된 패키지 scatterplot3d를 로드중입니다
> library("survival")
> library("coin")
> data("glioma", package="coin")
> library("survival")
> dim(glioma)
[1] 37  7
> glioma[1:10,]
   no. age    sex histology group event time
1    1  41 Female    Grade3   RIT  TRUE   53
2    2  45 Female    Grade3   RIT FALSE   28
3    3  48   Male    Grade3   RIT FALSE   69
4    4  54   Male    Grade3   RIT FALSE   58
5    5  40 Female    Grade3   RIT FALSE   54
6    6  31   Male    Grade3   RIT  TRUE   25
7    7  53   Male    Grade3   RIT FALSE   51
8    8  49   Male    Grade3   RIT FALSE   61
9    9  36   Male    Grade3   RIT FALSE   57
10  10  52   Male    Grade3   RIT FALSE   57
```

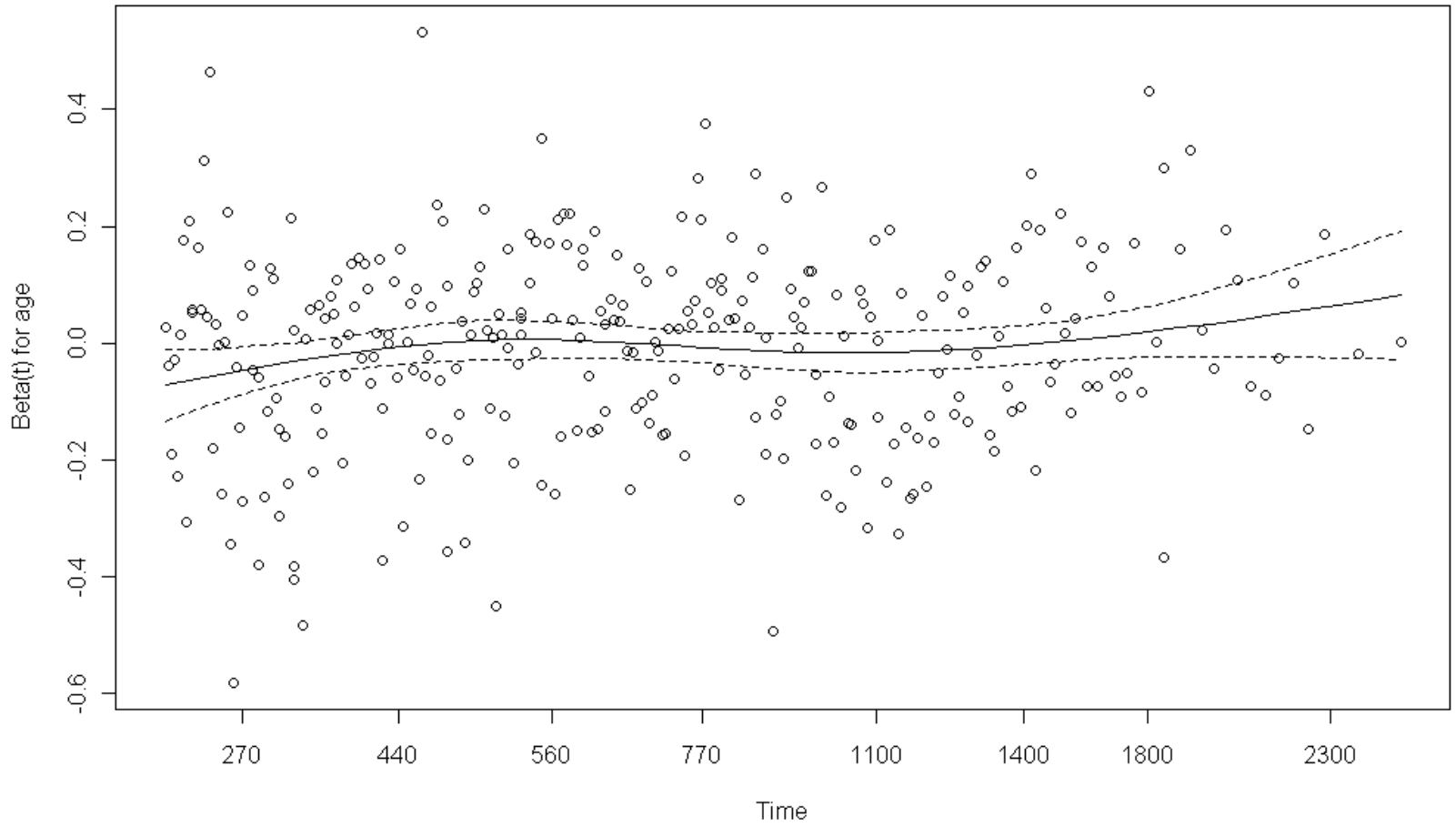
```
> g3 <- subset(glioma, histology == "Grade3")
> layout(matrix(1:2, ncol = 2))
> plot(survfit(Surv(time, event) ~ group, data = g3),
+   main = "Grade III Glioma", lty = c(2, 1),
+   ylab = "Probability", xlab = "Survival Time in Month",
+   legend.bty = "y", legend.text = c("Control", "Treated")
+   )
> # legend.bty  =  legend box
>
> g4 <- subset(glioma, histology == "GBM")
> plot(survfit(Surv(time, event) ~ group, data = g4),
+   main = "Grade IV Glioma", ylab = "Probability",
+   lty = c(2, 1), xlab = "Survival Time in Month",
+   xlim = c(0, max(glioma$time) * 1.05))
```


Grade III Glioma
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
Survival Time in Month
Control
Treated
Grade IV Glioma
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
Survival Time in Month

```
> library("ipred")
요구된 패키지 rpart를 로드중입니다
요구된 패키지 mlbench를 로드중입니다
요구된 패키지 nnet를 로드중입니다
요구된 패키지 class를 로드중입니다
> data("GBSG2", package = "ipred")
> GBSG2[1:10,]; dim(GBSG2)
   horTh age menostat tsize tgrade pnodes progrec estrec time cens
1     no  70      Post    21     II      3       48      66 1814    1
2    yes  56      Post    12     II      7       61      77 2018    1
3    yes  58      Post    35     II      9       52     271  712    1
4    yes  59      Post    17     II      4       60      29 1807    1
5     no  73      Post    35     II      1       26      65  772    1
6     no  32       Pre    57    III     24        0      13  448    1
7    yes  59      Post     8     II      2      181       0 2172    0
8     no  65      Post    16     II      1      192      25 2161    0
9     no  80      Post    39     II     30        0      59  471    1
10    no  66      Post    18     II      7        0       3 2014    0
[1] 686  10
> layout(matrix(1:1, ncol = 1))
> plot(survfit(Surv(time, cens) ~ horTh, data = GBSG2),
+   lty = 1:2, mark.time = FALSE, ylab = "Probability",
+   xlab = "Survival Time in Days")
> legend(250, 0.2, legend = c("yes", "no"), lty = c(2, 1),
+   title = "Hormonal Therapy", bty = "n")
```

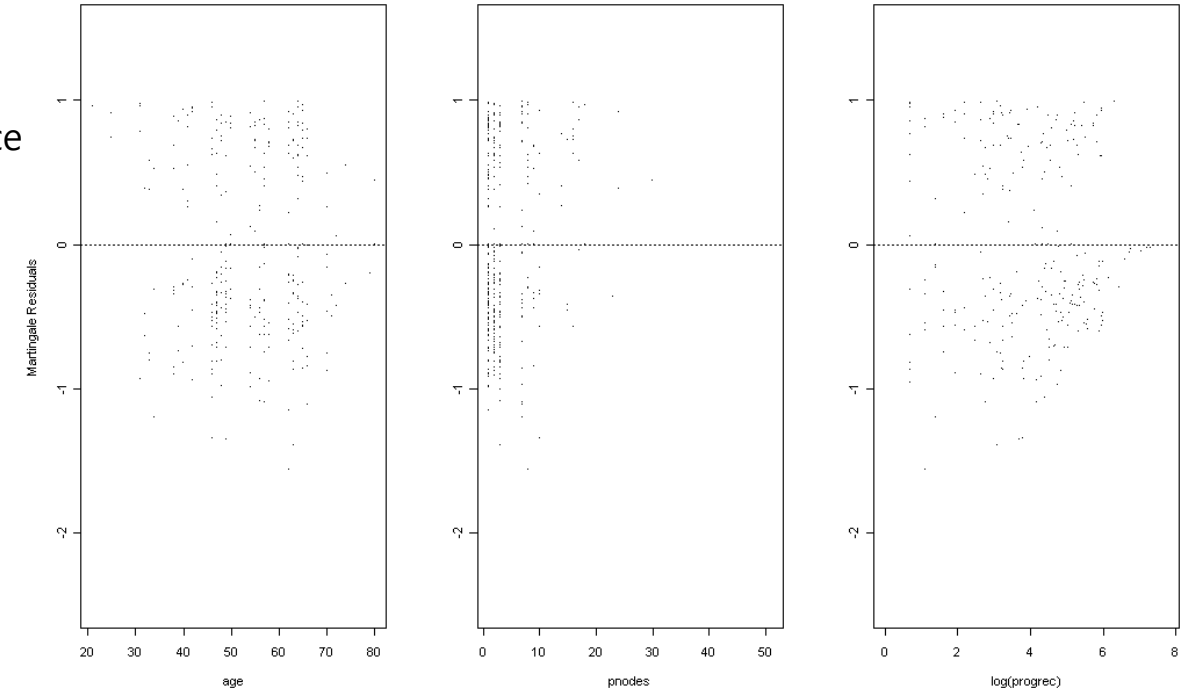

Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
Survival Time in Days
Hormonal Therapy
yes
no

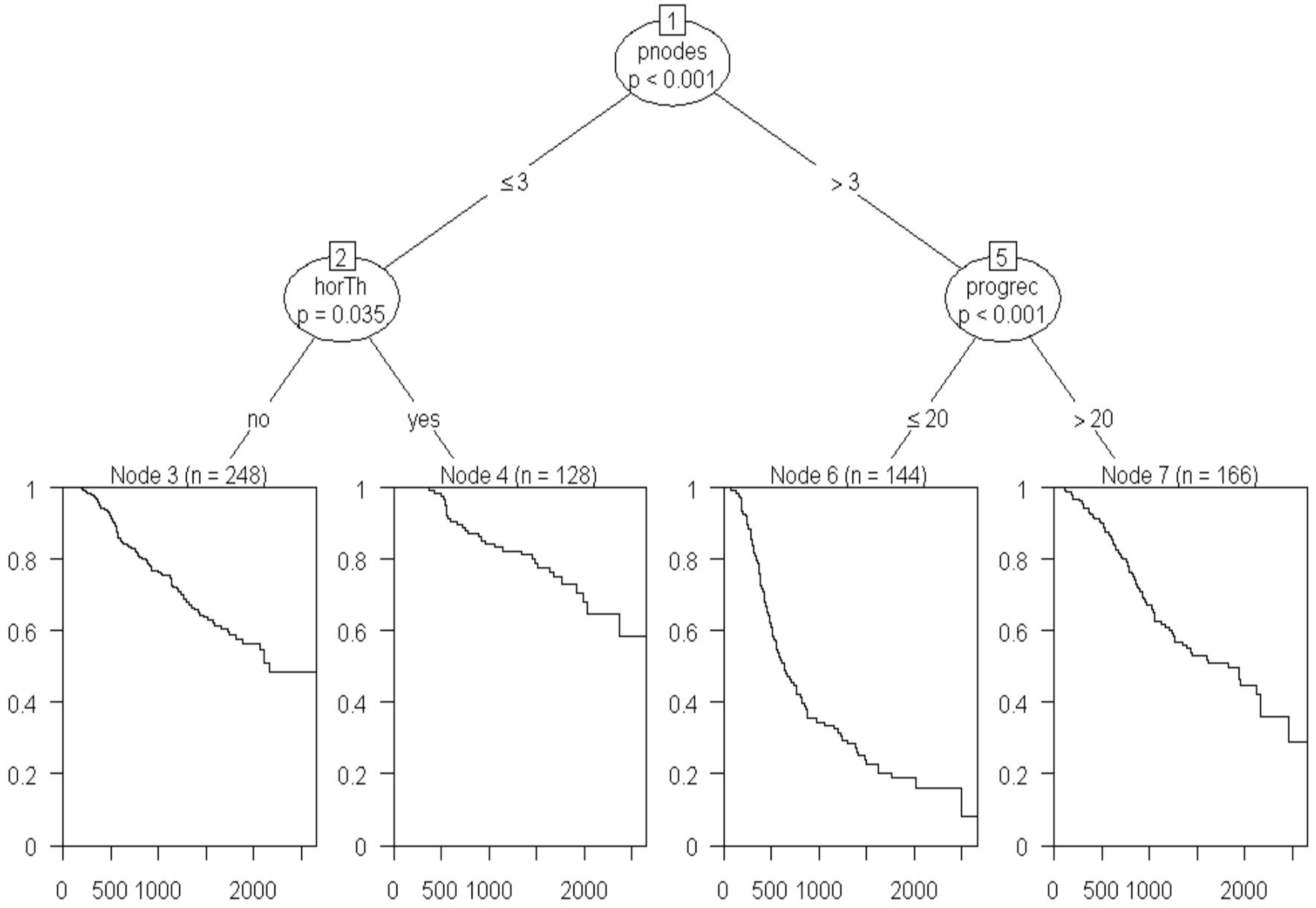
```
> GBSG2_coxph <- coxph(Surv(time, cens) ~ ., data = GBSG2)
> summary(GBSG2_coxph)
Call:
coxph(formula = Surv(time, cens) ~ ., data = GBSG2)

  n= 686
                   coef exp(coef) se(coef)      z        p
horThyes       -0.346278     0.707 0.129075 -2.683 7.3e-03
age            -0.009459     0.991 0.009301 -1.017 3.1e-01
menostatPost    0.258445     1.295 0.183476  1.409 1.6e-01
tsize           0.007796     1.008 0.003939  1.979 4.8e-02
tgrade.L        0.551299     1.736 0.189844  2.904 3.7e-03
tgrade.Q       -0.201091     0.818 0.121965 -1.649 9.9e-02
pnodes          0.048789     1.050 0.007447  6.551 5.7e-11
progrec        -0.002217     0.998 0.000574 -3.866 1.1e-04
estrec          0.000197     1.000 0.000450  0.438 6.6e-01

               exp(coef) exp(-coef) lower .95 upper .95
horThyes           0.707      1.414     0.549     0.911
age                0.991      1.010     0.973     1.009
menostatPost       1.295      0.772     0.904     1.855
tsize              1.008      0.992     1.000     1.016
tgrade.L           1.736      0.576     1.196     2.518
tgrade.Q           0.818      1.223     0.644     1.039
pnodes             1.050      0.952     1.035     1.065
progrec            0.998      1.002     0.997     0.999
estrec             1.000      1.000     0.999     1.001

Rsquare= 0.142   (max possible= 0.995 )
Likelihood ratio test= 105  on 9 df,   p=0
Wald test            = 115  on 9 df,   p=0
Score (logrank) test = 121  on 9 df,   p=0

> ci <- confint(GBSG2_coxph)
> exp(cbind(coef(GBSG2_coxph), ci))["horThyes",]
              2.5 %     97.5 %
0.7073155 0.5492178 0.9109233
>> # proportional assumption check #
> GBSG2_zph <- cox.zph(GBSG2_coxph)
> GBSG2_zph
                    rho     chisq        p
horThyes      -2.54e-02 1.96e-01 0.65778
age            9.40e-02 2.96e+00 0.08552
menostatPost -1.19e-05 3.75e-08 0.99985
tsize         -2.50e-02 1.88e-01 0.66436
tgrade.L      -1.30e-01 4.85e+00 0.02772
tgrade.Q       3.22e-03 3.14e-03 0.95530
pnodes         5.84e-02 5.98e-01 0.43941
progrec        5.65e-02 1.20e+00 0.27351
estrec         5.46e-02 1.03e+00 0.30967
GLOBAL               NA 2.27e+01 0.00695
> # some evidence of time varying effects for age and tumor grading
> plot(GBSG2_zph, var = "age")
```


Beta(t) for age
-0.6
-0.4
-0.2
0.0
0.2
0.4
270
440
560
770
1100
1400
1800
2300
Time

```
> layout(matrix(1:3, ncol = 3))
>  res <- residuals(GBSG2_coxph)
> plot(res ~ age, data = GBSG2, ylim = c(-2.5, 1.5),
+  pch = ".", ylab = "Martingale Residuals")
> abline(h = 0, lty = 3)
> plot(res ~ pnodes, data = GBSG2, ylim = c(-2.5, 1.5),
+  pch = ".", ylab = "")
> abline(h = 0, lty = 3)
> plot(res ~ log(progrec), data = GBSG2, ylim = c(-2.5, 1.5),
+  pch = ".", ylab = "")
> abline(h = 0, lty = 3)
```

```
library("party")
GBSG2_ctree <- ctree(Surv(time, ce
layout(matrix(1:1, ncol = 1))
plot(GBSG2_ctree)
```




1
pnodes
p < 0.001
≤ 3
> 3
2
horTh
p = 0.035
5
progrec
p < 0.001
no
yes
≤ 20
> 20
Node 3 (n = 248)
Node 4 (n = 128)
Node 6 (n = 144)
Node 7 (n = 166)
1
0.8
0.6
0.4
0.2
0
0
500
1000
2000

생존자료분석 요약

1. 생존율 산출
    1. 생명표법: 표본수가 많을 때 (>50)
    2. Kaplan-Meier method: 표본수가 적을 때 (< 50)
2. 생존율비교Mantel-Haenszel method
    1. Mantel-Haenszel method
    2. Log-rank method
    3. Gehan's generalized Wilcoxon
3. 생존기간에 영향을 주는 인자에 대한 HR 추정

    Cox proportional Hazard model

1. 단변량 분석 (평균, 표준편차, 비율 등 계산)
2. t-test , chi-square test (두 집단 비교)
3. 회귀분석, 로지스틱회귀분석 (다른 변인들의 효과를 보정한 후의 주변 수효과)

김호
서울대학교 보건대학원
hokim@snu.ac.kr

http:// plaza.snu. ac.kr/ ~hokim
-> 열린강의실 -> 수업외 자료방 ->
의학연구자료의 생존분석법