# 간경변증 발생 위험군 분류를 위한 SPAN의 유용성 평가

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

유 영 애

# 간경변증 발생 위험군 분류를 위한 SPAN의 유용성 평가

지도 안 상 훈 교수

이 논문을 석사 학위논문으로 제출함

2007년 12월 일

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

유 영 애

유영애의 석사 학위논문을 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

연세대학교 대학원

2007년 12월 일

# 차 례

# 표 차 례



# 그 림 차 례

## 국 문 요 약

# 간경변증 발생 위험군 분류를 위한 SPAN의 유용성 평가


임상의학분야에서 질병 발생의 예측이나 위험 요인을 분석하기 위해 통계학적 방법을 이용해왔다. 회귀분석방법들과 나무모형을 이용한 분류 분석 방법이 주로 이용되고 있다. 하지만 이런 방법들은 제한점을 가지고 있다. 이에 대한 대안으로 위험요인을 가진 하위 그룹을 찾는 알고리즘인 SPAN이 제안되었다. SPAN은 모든 가능한 변수의 조합 중에서 최상의 조합을 찾아내는 것으로, 종속변수와의 관계가 명백한 변수들의 조합만을 찾기 때문에 임상의학적으로 의미 있는 결론을 얻을 수 있으며, 이에 대한 해석과 적용이 쉽다는 장점을 가지고 있다.

본 논문에서는 SPAN을 이용하여 질병 발생 위험군 예측에 대한 유용성을 평가하고자 1994년부터 2005년까지 건강검진센터에서 건강검진을 받은 검진자 중 병원에 내원하여 간경변증 발생 여부에 대한 진단을 받은 4,093명의 검진자료를 이용하였다. 평가를 위해 로지스틱 회귀분석, 다항수준 회귀분석, QUEST를 가지고 민감도, 특이도, 정확도등의 측도를 통하여 비교 하였다.

SPAN에 의해 선택된 간경변증 발생 위험 인자는 B형간염 바이러스, C형 간염 항체, 가족력, 혈소판, 알파-태아단백으로 기존에 알려진 간경변 발생 위험인자를 잘 나타고 있다. 그리고 다른 분석 방법들에 비해 SPAN의 민감도가 높아 간경변증 발생 위험군 분류에 유용한 것으로 판단된다.


---


핵심되는 말 : SPAN, 로지스틱회귀분석, 다항수준회귀분석, QUEST, 민감도, 특이도, 정확도

# 제 1장 서론

## 1.1 연구 배경

임상의학분야에서 질병 발생의 예측이나 위험 요인을 분석하기 위해 통계학적 방법을 이용해왔다. 대표적으로 회귀분석방법 중에서 모수적 방법인 로지스틱 회귀분석이 널리 사용되어 왔다(Zhang H, 1996). 최근에는 회귀분석방법 중에서 비모수적 방법인 multivariate adaptive regression splines(이하 MARS)와 이를 응용한 다항수준 회귀분석(polychotomous regression)도 사용되고 있다. 이런 회귀분석방법의 장점은 질병발생에 대한 각 변수들의 중요도를 판단할 수 있고 아울러 예측이 용이하다는 것이다(정혜원, 2004).

회귀분석방법들과 함께 데이터마이닝의 한 분야로 나무모형을 이용한 분류(classification) 분석 방법 또한 빈번하게 이용되고 있는데, 나무모형을 이용한 분석 방법은 자료의 분포에 대한 특별한 가정이 필요 없고, 결과 해석이 쉽다는 장점을 가지고 있다(Austin PC, 2007). 이런 나무모형 중에 가장 대표적인 방법으로 classification and regression tree(이하 CART)가 있다. CART는 반복적인 탐색을 통해 최상의 분리점을 찾는 장점이 있다. 하지만 독립변수의 수와 그 범주가 많아지면 계산의 양이 많아져 시간이 오래 걸린다는 단점이 있다(Loh WY and Shih YS, 1997). 그래서 이를 보완하여 변수 선택은 통계적 유의성 검정을 사용하고, 분리점 선택은 탐색적 방법을 이용한 quick unbiased efficient statistical trees(이하 QUEST)도 사용되고 있다. 하지만 CART나 QUEST 같은 경우에는 위계적(hierarchical)으로 자료를 탐색하기 때문에 예상치 못한 결과가 종종 발생되어 해석을 난해하게 한다는 단점이 있다(Marshall RJ, 1995).

이런 단점을 보완하고자 비위계적(non-hierarchical) 분류분석방법인 search partition analysis(이하 SPAN)가 최근에 제안되었다. SPAN은 모든 가능한 변수의 조합 중에서 최상의 조합을 찾아내는 것으로, 종속변수와의 관계가 명백한 변수들의 조합만을 찾기 때문에 임상의학적으로 의미 있는 결론을 얻을 수 있으며, 이에 대한

해석과 적용이 쉽다는 장점을 가지고 있다(Marshall RJ, 2001).

## 1.2 연구 목적 및 내용

본 논문에서는 SPAN의 질병 발생 위험군 예측에 대한 유용성을 로지스틱 회귀분석, 다항수준 회귀분석, QUEST의 분석결과와 비교하며 평가하였다.

이 연구를 위해 1994년부터 2005년까지 건강검진센터에서 건강검진을 받은 검진자 중 병원에 내원하여 간경변증 발생 여부에 대한 진단을 받은 4,093명의 검진자료를 이용하는데, 자료를 둘로 나누어 하나는 훈련용 자료(training data)로 모형 설정을 위해 쓰고, 나머지 하나는 검증용 자료(test data)로 만들어진 모형의 검증을 위해 사용하였다.

논문의 구성은 제 1장 서론에서는 연구의 배경과 목적 및 내용에 대해서 언급하였고, 제 2장에서는 비교하고자 하는 연구 방법들을 소개하고, 제 3장에서는 SPAN의 이론적 배경을 설명한다. 그리고 제 4장에서는 실제 자료를 가지고 분석한 결과와 그에 대한 설명과 각 방법들의 비교에 대해서 언급하고, 마지막으로 제 5장에서는 네 가지 방법들에 대한 결론 및 고찰에 대해서 논의한다.

# 제 2장 분류 방법

## 2.1 로지스틱 회귀분석

로지스틱 회귀분석은 선형회귀분석과는 다르게 종속변수가 두 개의 범주로 이루어진 경우에 독립변수와의 관계를 살펴보기 위해 사용된다. 일반적으로 종속변수가 취할 수 있는 값은 어떤 사건이 발생된 경우를 1로 하고, 발생되지 않은 경우를 0으로 하여 독립변수가 주어졌을 때 사건발생의 조건부 확률을 로짓 변환하여 사용한다.

### 2.1.1 로지스틱 회귀모형

종속변수는 어떤 사건이 발생한 경우를 1, 발생되지 않은 경우를 0으로 표시한 이분형 변수이고, 독립변수는 연속형과 범주형 변수를 합쳐 $p$개가 있다고 하자. 이 때 $p$개의 독립변수에 대해 종속변수가 1을 가질 확률을 $P(Y=1|x_1,x_2,\cdots,x_p)$라고 하면, 이를 로짓 변환(logit transformation)을 통해 나타내는 것이 로지스틱 회귀모형으로 식으로 표현하면

$$\ln\left[\frac{P(Y=1|x_1,x_2,\cdots,x_p)}{1-P(Y=1|x_1,x_2,\cdots,x_p)}\right]=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p$$

이다. 이 식을 $P(Y=1|x_1,x_2,\cdots,x_p)$에 관하여 정리하면

$$P(Y=1|x_1,x_2,\cdots,x_p)=\frac{\exp(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}{1+\exp(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}$$

이고, 이를 로지스틱 반응함수라고 한다.

### 2.1.2 로지스틱 회귀계수의 추정과 검정

로지스틱 회귀모형에서 회귀계수를 추정하기 위해 최대우도법(maximum likelihood estimation)을 이용한다. 최대우도법은 우도함수를 최대로 하는 회귀계수를 추정하는 것이다. 우도함수의 최대값은 미분을 통해 얻게 되는데, 회귀계수의 최대우도 추정값은 비선형이므로 피셔의 스코어링방법(Fisher's method of scoring)이나 뉴튼-랩슨 방법(Newton-Raphson method)등과 같은 반복적인 추정방법에 의하여 근사값를 구한다(성웅현, 2001). 회귀계수를 추정한 후에 로지스틱 회귀모형에 대한 검정은 우도비 검정(likelihood ratio test), 왈드 검정(Wald test), 스코어 검정(score test) 등을 이용한다.

### 2.1.3 변수선택 방법

독립변수가 많은 경우에는 변수선택법을 이용하여 로지스틱 회귀모형의 독립변수를 선택 할 수 있다. 변수를 선택하는 방법에는 설명변수의 각각의 기여도에 따라 단계별로 하나씩 추가하면서 변수를 선택하는 전진선택법(forward selection)과 모든 독립변수를 포함한 완전모형에서 불필요한 변수를 단계별로 하나씩 제거해 나가는 후진제거법(backward elimination), 각 단계에서 변수의 선택과 제거를 반복하면서 독립변수를 결정하는 단계적 선택법(stepwise selection)이 있다.

### 2.1.4 사후 확률을 이용한 분류

로지스틱 회귀모형이 완성되면, 각 개체의 사후확률(posterior probability)을 이용하여 분류할 수 있다. 적절한 경계값(cut-off value)을 정하여 이 값을 기준으로 사후확률이 경계값보다 크면 집단 1로 분류하고, 경계값보다 작으면 집단 0으로 분류한다.

## 2.2 다항수준 회귀분석

### 2.2.1 다항수준 회귀모형

다항수준회귀분석은 MARS에 스플라인(spline) 함수를 적용시킨 것으로 모델은

$$\theta(k|\mathbf{x}) = \theta(k|\mathbf{x};\beta) = \sum_{j=1}^{p} \beta_{jk} B_j(\mathbf{x})$$

와 같이 표현된다. 여기서 $\theta(k|\mathbf{x})$는 스플라인 함수이고, $B_j(\mathbf{x})$는 기저 함수(basis function)로

$$(x-t)_+ = \begin{cases} x-t, \text{ if } x > t \\ 0, \quad otherwise' \end{cases}$$
$$(t-x)_+ = \begin{cases} t-x, \text{ if } x < t \\ 0, \quad otherwise \end{cases}$$

의 쌍으로 존재한다. 여기에서의 $t$는 독립변수 $X_j$의 $i$번째 관측치로 $x_{ij}$중 하나이다. 그러므로 기저함수의 집합

$$C = \{(X_j - t)_+, (t - X_j)_+\}_{\substack{t \in \{x_{1j}, x_{2j}, \cdots, x_{Nj}\} \\ j = 1,2,\cdots,p}}$$

이고, 만약 독립변수들이 각각 다른 값을 가진다면 $2Np$개의 기저함수가 있는 것이다.

### 2.2.2 모형 선택

다항수준 회귀모형은 $B_j(\mathbf{x})=1$에서 시작한다. 여기에 기저함수 한 개를 추가하고 최대우도법으로 $\beta$를 추정한다. 다음 또 다른 기저함수를 추가하고 $\beta$를 추정한다. 전과 후의 모형 중에 자료를 잘 설명하는 모형을 선택한다.

모형의 선택은 $AIC$(Akaike's information criterion)를 이용한다. $\hat{l}_v$을 $v$번째 모형의 적합된 로그우도(fitted log-likelihood)라고 하자. 그리고 이때의 $AIC$는

$$AIC_{\alpha,v}=-2\hat{l}_v+\alpha(K-1)p_v$$

이고, 여기서 $\alpha$는 $\log N$이다. 모든 모형 중에서 $AIC$가 가장 작은 모형을 선택한다.

## 2.3 Quick unbiased efficient statistical trees

### 2.3.1 이론적 배경

나무 모형 중에서 대표적으로 사용되는 CART의 경우 독립변수의 수와 그 범주의 개수가 많아지면 계산해야 하는 경우의 수가 지수적으로 증가하여 시간이 오래 걸린다(Loh WY and Shih YS, 1997). 이런 단점을 보완하고자 변수 선택과 분리점 선택을 두 단계로 나누어서 실시하는 알고리즘인 QUEST를 1997년 Loh와 Shih이 제안하였다.

### 2.3.2 변수 선택

첫 번째 단계인 변수 선택 단계는 종속변수에 따라 가장 유의한 독립변수를 찾는 과정이다. 독립변수가 연속형 변수일 경우에는 분산분석을 실시하고, 독립변수가 범주형 변수일 경우 카이제곱검정을 실시하여 유의확률을 얻는다. 각 독립변수에서 계산된 유의확률 중에 가장 작은 값을 가지는 독립변수를 선택한다.

### 2.3.3 분리점 선택

두 번째 단계는 분리점 선택 단계이다. 분리점을 선택하기에 앞서 첫 번째 단계에서 선택된 독립변수가 범주형 변수일 경우에는 CRIMCOORDs 변환을 통해 변수 변환을 한다. CRIMCOORDs 변환은 범주형 변수를 연속형 변수로 변환하는 것으로 주어진 범주형 변수의 주성분을 이용하여 차원 축소하여 새로운 변수를 얻는 방법이다. 범주형 변수의 CRIMCOORDs 변환 후, 선택된 독립변수를 가지고 이차판별분석이나 모든 가능한 분리점 중에 하나를 선택하는 CART의 분리점 선택 방법을 이용하여 분리점을 선택한다.

### 2.3.4 모형 선택

이 두 단계를 반복적으로 실시하면서 나무 모형을 만든다. 각 단계마다 오분류비용(misclassification cost)을 계산한다. 오분류 비용이 정해놓은 값보다 작은 나무모형을 최종모형으로 한다.

# 제 3장 Search partition analysis

## 3.1 이론적 배경

임상의학분야에서 위험군을 분류하고자 로지스틱 회귀분석이나 포아송 회귀분석을 이용한 모델링 방법을 이용해왔다. 최근에는 나무모형을 이용한 통계적 방법도 이용되고 있다. 나무모형을 이용한 방법은 자료를 세분화하는 과정에서 동질적인 하위그룹을 얻는 방법이다. 하지만 나무모형을 이용한 방법은 위계적인 것으로 결과를 적용하는데 제한점이 있다(Marshall RJ, 2001). 이에 대한 대안으로 SPAN이 제안 되었다.

## 3.2 속성

SPAN은 자료를 두개의 그룹으로 나누는 알고리즘이다. 여기서 두 개의 그룹을 $S$와 $S'$라 하자. 둘로 나뉜 분류의 결과는 속성들의 부울 결합(Boolean combinations)을 이용하여 표현된다. 여기서 말하는 속성이란 독립변수로부터 얻은 특징으로, 예를 들면 '나이가 40세 이상이다', '간경변증에 대한 가족력이 있다' 라는 것으로 질병에 대한 위험인자로 생각할 수 있는 것들이다. 속성들의 부울 결합은 자료를 두 개의 공간으로 나누게 되는데, 이때의 목표는 $S$, $S'$ 각각을 가장 동질적으로 만드는 분류 결과 $A$, $A'$을 찾는 것이다.

속성은 이분형의 형태로만 가능하다. 그래서 독립변수가 이분형 변수일 경우에는 문제가 되지 않지만, 독립변수가 범주형이나 연속형 변수일 때에는 연구자가 기존에 알려져 있는 정보를 이용하거나 탐색적 자료 분석을 통해 속성을 정의한다. 속성 $X$는 연속형 변수 $x$에 대해서 임의의 경계값 $C$를 기준으로

$$X = \{x > C\}$$

로 나타내고, 범주형 변수에 대해서는 임의의 범주 $C$에 속하는 것을 속성으로 정의하고 이는

$$X = \{x \in C\}$$

로 나타낸다. 속성 정의를 한 후에 $S$를 잘 반영하는 $m$개의 속성을 선택하여 속성들의 집합인

$$T_m = \{X_1, X_2, \cdots, X_m\}$$

을 정한다. 이때의 $m$의 크기는 연구자가 임의로 선택한다. 보통 $m$은 카이제곱 통계량을 이용하여 통계적으로 유의한 속성의 수로 결정한다.

## 3.3 부울 결합

$T_m$의 크기인 $m$을 정하고 나면, 부울 결합의 크기를 정한다. 부울 결합의 일반적인 형태는

$$A = K_1 \cup K_2 \cup \cdots \cup K_q$$

이고, 여기서 $K_i$는 $T_m$의 원소들 중에서 $p_i$개 원소들의 결합이다. $A$가 결정되면 나머지 부울 결합은

$$A' = J_1 \cup J_2 \cup \cdots \cup J_{q'}$$

로 표현되고, 여기서 $J_i$는 $K_i$를 구성하는 원소들의 다른 한쪽들의 결합이다. $m$과 $q$, $p_i$가 주어졌을 때 부울 결합의 모든 가능한 수를 [표 1]에 정리해 보았다. $m$, $q$, $p_i$가 늘어남에 따라 가능한 부울 결합수가 지수적으로 늘어남을 확인할 수 있다.

표 1. $m$, $q$, $p_i$에 따른 부울 결합의 수

| $p_1,\cdots,p_q$ | $m$=8 | $m$=12 | $m$=16 | $m$=20 |
|---|---|---|---|---|
| 1 | 8 | 12 | 16 | 20 |
| 2 | 28 | 66 | 120 | 190 |
| 3 | 56 | 220 | 560 | 1,140 |
| 1,1 | 28 | 66 | 120 | 190 |
| 2,1 | 168 | 660 | 1,680 | 3,420 |
| 3,1 | 280 | 1,980 | 7,280 | 19,380 |
| 2,2 | 378 | 2,145 | 7,140 | 17,955 |
| 3,2 | 1,400 | 13,860 | 65,520 | 213,180 |
| 3,3 | 1,540 | 24,090 | 156,520 | 649,230 |
| 1,1,1 | 56 | 220 | 560 | 1,140 |
| 2,1,1 | 440 | 2,970 | 10,840 | 29,070 |
| 2,2,1 | 1,680 | 17,820 | 87,360 | 290,770 |
| 2,2,2 | 3,276 | 45,760 | 280,840 | 1,125,180 |

연구자에 의해 정해진 $m$, $q$, $p_i$에 대한 부울 결합들 중에 $S$, $S'$를 각각을 가장 동질적으로 하는 부울 결합을 찾는 것은 불순도의 기준인

$$i(P) = -P\log P - (1-P)\log(1-P)$$

를 사용한

$$G = i(P_S) - P_A i(P_{S|A}) - P_{A'} i(P_{S|A'})$$

를 이용한다. 여기서 $P_A$와 $P_{A'}$는 $A$와 $A'$의 확률이고, $P_{S|A}$와 $P_{S|A'}$는 $A$와 $A'$가 주어졌을 때 $S$의 조건부 확률이다. 정해진 $m$, $q$, $p_i$에 대한 모든 가능한 경우의 수만큼의 부울 결합에 대해 $G$값을 모두 구해 이를 가장 크게 하는 부울 결합을 찾는다.

## 3.4 모형 크기의 제한

최상의 부울 결합을 찾는 과정에서 그 크기가 무한히 커지는 것을 막기 위해

$$G - c\beta$$

를 이용하여 제한을 주는데, 이때의

$$c = q + q' - 1$$

이고 $\beta$는 $c$가 하나씩 늘어남에 따라 나타나는 기울기로 이때 기울기의 증가하는 정도가 작아지면 그 전단계의 $c$에서 $G$를 최대로 하는 부울 결합을 선택한다.

# 제 4장 간경변 발생 위험군 분류 결과

## 4.1 자료에 대한 개요

건강검진 센터에서 1994년 5월부터 2005년 9월 사이에 건강검진을 받은 총 124,121명의 건강검진 자료를 바탕으로 중복된 검진자들은 가장 최근의 정보를 담아 총 85,458명을 추출하였다. 이들 중 다시 병원에 내원하여 소화기내과 검진을 받은 8,031명 중에서 2000년부터 실시한 문진항목을 추가하고, 결측치가 존재하는 대상은 제외하여 최종적으로 4,093명을 분석 자료로 사용하였다. 4,093명 중 간경변 발생분포를 살펴보면 간경변 발생자는 501명이고, 비발생자는 3,592명이다. 본 논문에서는 간경변 발생 위험군을 분류하고, 분류된 결과를 검증하기 위하여 4,093명 자료를 1:1의 비율로 임의로 나누어 분석하였다.(Picard RR and Berk KN, 1990) 훈련용 자료에서의 간경변 발생자는 251명이고, 검증용 자료에서의 간경변 발생자는 250명으로 [표 2]와 같다.

표 2. 간경변 발생 분포

| | 전체 자료 | 훈련용 자료 | 검증용 자료 |
|---|---|---|---|
| 간경병 비발생 | 3,592 (87.76%) | 1,795 (87.73%) | 1,797 (87.79%) |
| 간경변 발생 | 501 (12.24%) | 251 (12.27%) | 250 (12.21%) |
| 계 | 4,093 (100.0%) | 2,046 (100.0%) | 2,047 (100.0%) |

## 4.2 분석에 쓰인 건강검진 항목

분석을 위해 기존에 알려진 간경변 발생 위험인자를 바탕으로 건강검진 항목에서 기초 정보, 혈액 검사, 간기능 검사, 혈청지질 검사, 종양혈청 검사, 대사 및 전해질 검사, 간염 검사, 뇨 검사, 문진 항목을 사용하였다. [표 3]은 검진항목의 세부내용을

정리한 것이다.

표 3. 건강검진 세부항목

| 검사항목 | 변수 | 영문 | 한글 |
|---|---|---|---|
| 기초정보 | sex | | 성별 |
| | age | | 연령 |
| 혈액검사 | RBC | red blood cell | 적혈구 |
| | Hb | hemoglobin | 헤모글로빈 |
| | Hct | hematocrit | 헤마토크리트 |
| | MCV | mean corpuscular volume | 평균적혈구용적 |
| | MCH | mean corpuscular hemoglobin | 평균적혈구혈색소량 |
| | MCHC | mean corpuscular hemoglobin concentration | 평균적혈구혈색소농도 |
| | WBC | white blood cell | 백혈구 |
| | LYM | lymphocyte | 림프구 |
| | EOS | eosinocyte | 호산구 |
| | BAS | basophil leukocyte | 호염구 |
| | platelet | platelet | 혈소판 |
| 대사 및 전해질 | Na | sodium | 나트륨 |
| | K | potassium | 칼륨 |
| | Cl | chlorine | 염소 |
| | $CO_2$ | carbon dioxide | 이산화탄소 |
| | Ca | calcium | 칼슘 |
| | P | phosphorus | 인 |
| | glucose | blood glucose | 혈당 |
| | BUN | blood urea nitrogen | 혈중요소질소 |
| | creatinine | creatinine | 크레아티닌 |
| | uric acid | uric acid | 요산 |

표 3. 건강검진 세부항목(계속)

| 검사항목 | 변수 | 영문 | 한글 |
|---|---|---|---|
| 간기능검사 | T.protein | total protein | 총단백 |
| | albumin | albumin | 알부민 |
| | T.bilirubin | total bilirubin | 총빌리루빈 |
| | Alk.phos | alkaline phophatase | 알칼리성 포스파타제 |
| | AST | aspartate aminotransferase | 아스파르테이트 아미노전이효소 |
| | ALT | alanine aminotransferase | 알라닌아미노전이효소 |
| | γ-GT | gamma-glutamyl transferase | 감마-글루타밀전이효소 |
| | LDH | lactic dehydrogenase | 젖산탈수소효소 |
| 혈청지질 | T.cholesterol | total cholesterol | 총콜레스테롤 |
| | triglyceride | triglyceride | 중성지방 |
| | HDL | high-density lipoprotein | 고밀도콜레스테롤 |
| 간염검사 | HBsAg | hepatitis b virus | B형간염 바이러스 |
| | AntiHBc | | B형간염 C항체 |
| | AntiHCV | antihepatitis C virus | C형간염 항체 |
| 종양혈청 | α-FP | alpha-fetoprotein | 알파-태아단백 |
| | CEA | carcinoembryonic antigen | 태아성암항원 |
| 뇨검사 | SG | specific gravity | 비중 |
| | pH | hydrogen ion concentration | 수소이온농도지수 |
| | protein | protein | 단백질 |
| | urine glucose | urine glucose | 요당 |
| | ketone | ketone body | 케톤체 |
| | blood | occult blood | 잠혈 |
| | urobilinogen | urobilinogen | 우로빌리노겐 |
| | bilirubin | bilirubin | 빌리루빈 |
| | nitrite | nitrite | 아질산염 |
| | UWBC | white blood cell | 백혈구 |
| 문진 | family history | | 가족력 |
| | drinking | | 음주력 |
| | excercise | | 운동여부 |

## 4.3 전체 자료의 카이제곱 검정

4,093명의 전체 자료를 건강검진표의 기준을 기초로 하여 양성과 음성기준을 주어 51개의 변수를 두 개의 범주로 나누었다. 그 기준은 [표 4]에 정리하였다.

표 4. 변수의 분리 기준

| 변수 | 단위 | 양성기준 | 음성기준 |
|---|---|---|---|
| sex | | 남 | 여 |
| age | 세 | >=40 | <40 |
| RBC | $\times 10^6$/uL | 남 : <4.7 or >6.1<br>여 : <4.2 or >5.4 | 남 : 4.7-6.1<br>여 : 4.2-5.4 |
| Hb | g/dL | 남 : <14<br>여 : <12 | 남 : >=14<br>여 : >=12 |
| Hct | % | 남 : <42 or >52<br>여 : <37 or >47 | 남 : 42-52<br>여 : 37-47 |
| MCV | fL | <80 | >=80 |
| MCH | pg | <27 | >=27 |
| MCHC | g/dL | <33 | >=33 |
| WBC | $\times 10^6$/uL | >11 | <=11 |
| LYM | % | <19 or >48 | 19-48 |
| EOS | % | >7 | <=7 |
| BAS | % | >1.5 | <=1.5 |
| platelet | $10^3$/uL | <130 | >=130 |
| Na | mM/L | <135 or >145 | 135-145 |
| K | mM/L | <3.5 or >5.5 | 3.5-5.5 |
| Cl | mM/L | <98 or >110 | 98-110 |
| $CO_2$ | mM/L | >30 | <=30 |
| Ca | mg/dL | <8.8 or >11 | 8.8-11 |
| P | mg/dL | <2.5 or >4.5 | 2.5-4.5 |
| glucose | mg/dL | <70 or >110 | 70-110 |
| BUN | mg/dL | <5 or >25 | 5-25 |
| creatinine | mg/dL | >1.4 | <=1.4 |
| uric acid | mg/dL | >6 | <=6 |

표 4. 변수의 분리 기준 (계속)

| 변수 | 단위 | 양성기준 | 음성기준 |
|---|---|---|---|
| T.protein | g/dL | <6 | >=6 |
| albumin | g/dL | <3.3 or >5.3 | 3.3-5.3 |
| T.bilirubin | mg/dL | >=1.2 | <1.2 |
| Alk.phos | IU/L | >115 | <=115 |
| AST/ALT | IU/L | >1 | <=1 |
| γ-GT | IU/L | 남 : >55<br>여 : >35 | 남 : <=55<br>여 : <=35 |
| LDH | IU/L | >455 | <=455 |
| T.cholesterol | mg/dL | >=200 | <200 |
| triglyceride | mg/dL | <140 | >=140 |
| HDL | mg/dL | <30 | >=30 |
| HBsAg | | 양성 | 음성 |
| AntiHBc | | 양성 | 음성 |
| AntiHCV | | 양성 | 음성 |
| α-FP | IU/L | >20 | <=20 |
| CEA | ng/mL | >=5 | <5 |
| SG | | <1.003 or >1.030 | 1.003-1.030 |
| pH | | >8 | <=8 |
| protein | | 양성 | 음성 |
| urine glucose | | 양성 | 음성 |
| ketone | | 양성 | 음성 |
| blood | | 양성 | 음성 |
| urobilinogen | EU/dl | >=0.2 | <0.2 |
| bilirubin | | 양성 | 음성 |
| nitrite | | 양성 | 음성 |
| UWBC | | 양성 | 음성 |
| family history | | 유 | 무 |
| drinking | | 하루평균 소주 1병이상 | 그 외 |
| excercise | | 운동한다 | 운동하지 않는다 |

이분형으로 나눈 51개 변수에 대한 카이제곱검정 결과는 [표 5]과 같다. 결과를 보면 sex, LYM, platelet, K, Ca, albumin, T.bilirubin, Alk.phos, AST/ALT, rGT, LDH, T.cholesterol, triglyceride, HDL, HBsAg, AntiHBc, AntiHCV, $\alpha$FP, CEA, ketone, urobilinogen, bilirubin, family history, drinking에서 간경변 발생군과 비발생군 간에 유의한 차이가 있는 것으로 나타났다.

표 5. 독립변수의 카이제곱검정

| | 비간경변 (n=3,592) | | 간경변 (n=501) | | |
|---|---|---|---|---|---|
| 변수 | 빈도 | (%) | 빈도 | (%) | 유의확률 |
| sex | 1,919 | (53.42) | 329 | (65.67) | 0.0001 |
| age | 2,902 | (80.79) | 404 | (80.64) | 0.9356 |
| RBC | 1822 | (50.72) | 259 | (51.70) | 0.6833 |
| Hb | 571 | (15.90) | 67 | (13.37) | 0.1447 |
| Hct | 1,116 | (31.07) | 141 | (28.14) | 0.1836 |
| MCV | 63 | ( 1.75) | 6 | ( 1.20) | 0.3649 |
| MCH | 78 | ( 2.17) | 6 | ( 1.20) | 0.1498 |
| MCHC | 402 | (11.19) | 50 | ( 9.98) | 0.4177 |
| WBC | 65 | ( 1.81) | 8 | ( 1.60) | 0.7360 |
| LYM | 379 | (10.55) | 86 | (17.17) | 0.0001 |
| EOS | 42 | ( 1.17) | 5 | ( 1.00) | 0.7361 |
| BAS | 8 | ( 0.22) | 1 | ( 0.20) | 1.0000 |
| platelet | 52 | ( 1.45) | 85 | (16.97) | 0.0001 |
| Na | 114 | ( 3.17) | 13 | ( 2.59) | 0.4839 |
| K | 28 | ( 0.78) | 9 | ( 1.80) | 0.0385 |
| Cl | 67 | ( 1.87) | 13 | ( 2.59) | 0.2691 |
| $CO_2$ | 120 | ( 3.34) | 18 | ( 3.59) | 0.7697 |
| Ca | 77 | ( 2.14) | 24 | ( 4.79) | 0.0003 |
| P | 132 | ( 3.67) | 25 | ( 4.99) | 0.1510 |
| glucose | 454 | (12.64) | 64 | (12.77) | 0.9320 |
| BUN | 29 | ( 0.81) | 7 | ( 1.40) | 0.1969 |
| creatinine | 125 | (3.48) | 10 | ( 2.00) | 0.0815 |
| uric acid | 831 | (23.13) | 114 | (22.75) | 0.8499 |

표 5. 독립변수의 카이제곱검정(계속)

| 변수 | 비간경변 (n=3,592) 빈도 | (%) | 간경변 (n=501) 빈도 | (%) | 유의확률 |
|---|---|---|---|---|---|
| T.protein | 13 | ( 0.36) | 3 | ( 0.60) | 0.4341 |
| albumin | 19 | ( 0.53) | 9 | ( 1.80) | 0.0013 |
| T.bilirubin | 513 | (14.28) | 103 | (20.56) | 0.0002 |
| Alk.phos | 187 | ( 5.21) | 74 | (14.77) | 0.0001 |
| AST/ALT | 1,858 | (51.73) | 228 | (45.51) | 0.0091 |
| r-GT | 714 | (19.88) | 184 | (36.73) | 0.0001 |
| LDH | 195 | ( 5.43) | 56 | (11.18) | 0.0001 |
| T.cholesterol | 1,444 | (40.20) | 148 | (29.54) | 0.0001 |
| triglyceride | 2,179 | (60.66) | 366 | (73.05) | 0.0001 |
| HDL | 48 | ( 1.34) | 17 | ( 3.39) | 0.0006 |
| HBsAg | 151 | ( 4.20) | 245 | (48.90) | 0.0001 |
| AntiHBc | 2,117 | (58.94) | 403 | (80.44) | 0.0001 |
| AntiHCV | 27 | ( 0.75) | 65 | (12.97) | 0.0001 |
| $\alpha$-FP | 1 | ( 0.03) | 33 | ( 6.59) | 0.0001 |
| CEA | 232 | ( 6.46) | 56 | (11.18) | 0.0001 |
| SG | 4 | ( 0.11) | 0 | ( 0.00) | 1.0000 |
| pH | 36 | ( 1.00) | 8 | ( 1.60) | 0.2267 |
| protein | 381 | (10.61) | 66 | (13.17) | 0.0844 |
| urineglucose | 101 | ( 2.81) | 15 | ( 2.99) | 0.8179 |
| ketone | 168 | ( 4.68) | 40 | ( 7.98) | 0.0016 |
| blood | 1,154 | (32.13) | 143 | (28.54) | 0.1063 |
| urobilinogen | 117 | ( 3.26) | 47 | ( 9.38) | 0.0001 |
| bilirubin | 112 | ( 3.12) | 48 | ( 9.58) | 0.0001 |
| nitrite | 25 | ( 0.70) | 4 | ( 0.80) | 0.7746 |
| UWBC | 435 | (12.11) | 67 | (13.37) | 0.4195 |
| family history | 198 | ( 5.51) | 243 | (48.50) | 0.0001 |
| drinking | 548 | (15.26) | 164 | (32.73) | 0.0001 |
| exercise | 1,968 | (54.79) | 296 | (59.08) | 0.0702 |

## 4.4 훈련용 자료의 카이제곱 검정

2,046명의 훈련용 자료에서 카이제곱검정 결과는 [표 6]과 같다. 이 중에서 유의한 차이를 보인 sex, LYM, platelet, T.bilirubin, Alk.phos, AST/ALT, rGT, LDH, T.cholesterol, triglyceride, HBsAg, AntiHBc, AntiHCV, $\alpha$FP, CEA, ketone, urobilinogen, bilirubin, family history, drinking 이상 20개의 변수를 분석에 이용하였다.

표 6. 훈련용 자료에서 독립변수의 카이제곱검정

| 변수 | 비간경변 (n=1,795) 빈도 | (%) | 간경변 (n=251) 빈도 | (%) | 유의확률 |
|---|---|---|---|---|---|
| sex | 919 | (51.20) | 165 | (65.74) | 0.0001 |
| age | 1,447 | (80.61) | 194 | (77.29) | 0.2160 |
| RBC | 902 | (50.25) | 133 | (52.99) | 0.4165 |
| Hb | 274 | (15.26) | 32 | (12.75) | 0.2952 |
| Hct | 552 | (30.75) | 72 | (28.69) | 0.5053 |
| MCV | 33 | ( 1.84) | 2 | ( 0.80) | 0.3051 |
| MCH | 41 | ( 2.28) | 3 | ( 1.20) | 0.2653 |
| MCHC | 207 | (11.53) | 19 | ( 7.57) | 0.0607 |
| WBC | 37 | ( 2.06) | 5 | ( 1.99) | 0.9422 |
| LYM | 186 | (10.36) | 54 | (21.51) | 0.0001 |
| EOS | 25 | ( 1.39) | 3 | ( 1.20) | 1.0000 |
| BAS | 6 | ( 0.33) | 1 | ( 0.40) | 0.6005 |
| platelet | 27 | ( 1.50) | 38 | (15.14) | 0.0001 |
| Na | 59 | ( 3.29) | 8 | ( 3.19) | 0.9338 |
| K | 13 | ( 0.72) | 5 | ( 1.99) | 0.0596 |
| Cl | 39 | ( 2.17) | 6 | ( 2.39) | 0.8256 |
| $CO_2$ | 54 | ( 3.01) | 7 | ( 2.79) | 0.8481 |
| Ca | 41 | ( 2.28) | 8 | ( 3.19) | 0.3807 |
| P | 66 | ( 3.68) | 10 | ( 3.98) | 0.8095 |
| glucose | 205 | (11.42) | 28 | (11.16) | 0.9014 |
| BUN | 15 | ( 0.84) | 3 | ( 1.20) | 0.4768 |
| creatinine | 64 | ( 3.57) | 4 | ( 1.59) | 0.1026 |
| uric acid | 403 | (22.45) | 57 | (22.71) | 0.9270 |

표 6. 훈련용 자료에서 독립변수의 카이제곱검정(계속)

| 변수 | 비간경변 (n=1,795) 빈도 | (%) | 간경변 (n=251) 빈도 | (%) | 유의확률 |
|---|---|---|---|---|---|
| T.protein | 6 | ( 0.33) | 1 | ( 0.40) | 0.6005 |
| albumin | 7 | ( 0.39) | 3 | ( 1.20) | 0.1139 |
| T.bilirubin | 251 | (13.98) | 53 | (21.12) | 0.0029 |
| Alk.phos | 81 | ( 4.51) | 41 | (16.33) | 0.0001 |
| AST/ALT | 947 | (52.76) | 107 | (42.63) | 0.0026 |
| r-GT | 343 | (19.11) | 97 | (38.65) | 0.0001 |
| LDH | 100 | ( 5.57) | 31 | (12.35) | 0.0001 |
| T.cholesterol | 702 | (39.11) | 74 | (29.48) | 0.0032 |
| triglyceride | 1,091 | (60.78) | 175 | (69.72) | 0.0063 |
| HDL | 24 | ( 1.34) | 6 | ( 2.39) | 0.2535 |
| HBsAg | 75 | ( 4.18) | 113 | (45.02) | 0.0001 |
| AntiHBc | 1,061 | (59.11) | 199 | (79.28) | 0.0001 |
| AntiHCV | 14 | ( 0.78) | 33 | (13.15) | 0.0001 |
| $\alpha$-FP | 0 | ( 0.00) | 12 | ( 4.78) | 0.0001 |
| CEA | 110 | ( 6.13) | 29 | (11.55) | 0.0014 |
| SG | 0 | ( 0.00) | 0 | ( 0.00) | |
| pH | 20 | ( 1.11) | 4 | ( 1.59) | 0.5255 |
| protein | 195 | (10.86) | 29 | (11.55) | 0.7429 |
| urineglucose | 47 | ( 2.62) | 10 | ( 3.98) | 0.2182 |
| ketone | 83 | ( 4.62) | 22 | ( 8.76) | 0.0054 |
| blood | 621 | (34.60) | 72 | (28.69) | 0.0638 |
| urobilinogen | 66 | ( 3.68) | 22 | ( 8.76) | 0.0002 |
| bilirubin | 61 | ( 3.40) | 25 | ( 9.96) | 0.0001 |
| nitrite | 19 | ( 1.06) | 0 | ( 0.00) | 0.1553 |
| UWBC | 234 | (13.04) | 36 | (14.34) | 0.5668 |
| family history | 99 | ( 5.52) | 122 | (48.61) | 0.0001 |
| drinking | 273 | (15.21) | 85 | (33.86) | 0.0001 |

## 4.5 변수선택

간경변 발생 위험군을 분류하기 위해 간경변의 위험인자를 찾고, 이를 통해 분류의 정확도를 파악하였다. 본 논문에서는 4,093명의 20개의 이분형 독립변수를 가지고 SPAN과 단계적 로지스틱 회귀분석, 다항수준 회귀분석, QUEST의 성능을 비교하였다.

분류 결과에 대한 평가를 위해서 민감도와 특이도, 위양성율과 위음성율, 정확도(total accuracy)를 사용하였다. 민감도(sensitivity)는 실제 간경변 발생자를 예측분류에서도 간경변 발생자로 분류하는 비율을 나타내는 것이고, 특이도(specificity)는 실제 간경변 비발생자를 예측분류에서도 간경변 비발생자로 분류하는 비율을 나타낸다. 위양성율(false positive)은 실제 간경변 비발생자를 간경변 발생자로 잘못 분류한 비율이고, 위음성율(false negative)은 실제 간경변 발생자를 간경변 비발생자로 잘못 분류한 비율이다.

본 논문에서는 간경변에 대한 위험군 분류에 목적이 있으므로 민감도를 높이는 것을 중점으로 보도록 한다. 그리고 각 방법들의 분류 능력을 평가하기 위해서 앞에서 언급했듯이 훈련용 자료와 검증용 자료로 나누어 살펴보았다.

### 4.5.1 로지스틱 회귀분석

단계적 로지스틱 회귀분석에서는 LYM, platelet, Alk.phos, HBsAg, AntiHCV, family history, drinking이 간경변 발생의 위험인자로 나타났다. 선택된 변수의 계수와 유의확률은 [표 7]과 같다.

표 7. 단계적 로지스틱 회귀분석에서의 변수 선택

| Variable | Coefficient | Standard error | Wald Chi-Square | P-value |
|---|---|---|---|---|
| LYM | 0.6276 | 0.2412 | 6.7736 | 0.0093 |
| platelet | 1.0963 | 0.3960 | 7.6648 | 0.0056 |
| Alk.phos | 1.2115 | 0.3051 | 15.7644 | <.0001 |
| HBsAg | 2.9003 | 0.2150 | 181.8968 | <.0001 |
| AntiHCV | 3.7193 | 0.3831 | 94.2768 | <.0001 |
| family history | 2.4572 | 0.2080 | 139.6128 | <.0001 |
| drinking | 1.1957 | 0.2057 | 33.7863 | <.0001 |

### 4.5.2 다항수준회귀분석

다항수준회귀분석에서는 Alk.phos, HBsAg, AntiHCV, family history, drinking이 간경변 발생 위험인자로 이루어진 모형이 선택되었다. 각 변수의 기저함수의 회귀계수는 [표 8]과 같다.

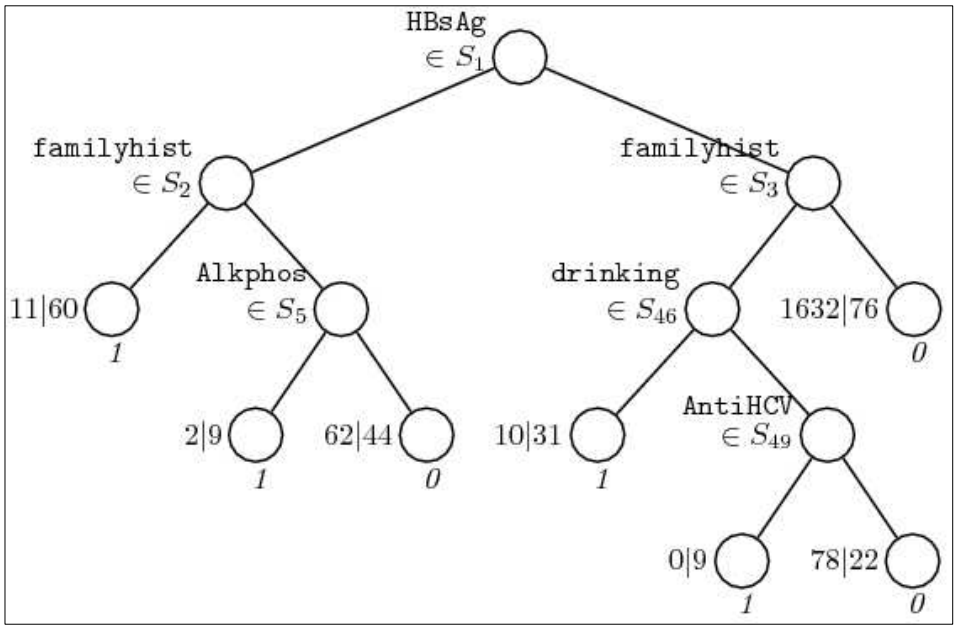
표 8. 다항수준회귀분석에서의 변수 선택

| | $(x-t)_+$ | $(t-x)_+$ |
|---|---|---|
| Alk.phos | −1.355 | 0 |
| HBsAg | −3.036 | 0 |
| AntiHCV | −3.766 | 0 |
| family history | −2.129 | 0 |
| drinking | −0.831 | 0 |

### 4.5.3 Quick unbiased efficient statistical trees

QUEST에서 사전확률은 자료의 발생확률을 계산하여 사용하고, 오분류비용은 같게 하고, 변수선택은 통계적 방법으로, 분리 기준은 지니지수로 하여 모형을 설정하였다. 간경변증 발생 위험인자는 Alk.phos, HBsAg, AntiHCV, family history, drinking으로 5개가 선택되었다. [그림 1]은 QUEST에서 선택된 변수들의 나무 모형으로 제일 왼쪽의 가지를 보면 HBsAg가 양성이고, family history가 있으면 간경변 위험군으로 분류되는 것을 보여준다.

그림 1. QUEST에서 선택된 변수들의 나무 모형



### 4.5.4 Search partition analysis

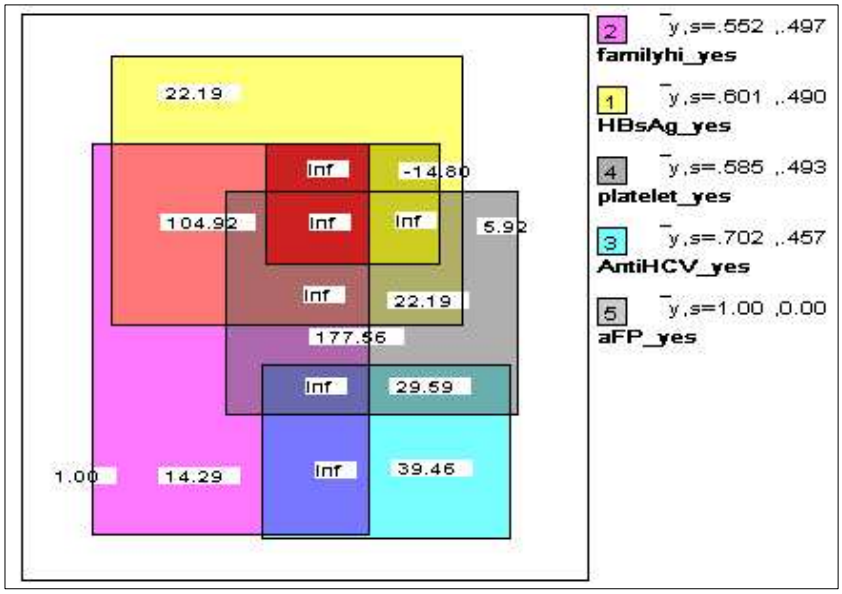
SPAN에서는 민감도와 특이도가 최대가 되는 지점인 변수 5개로 이루어진 모형을 선택하였다. 선택된 간경변증 발생 위험인자는 platelet, HBsAg, AntiHCV, $\alpha$-FP, family history로 이를 부울 결합을 이용하여 표현하면

간경변증 위험군 : B형간염 바이러스=양성 or 가족력=유 or
C형간염항체=양성 or 혈소판<130($10^3$/uL) or
알파태아성단백>20(IU/L)

간경변증 비위험군 : B형간염 바이러스=음성 and 가족력=무 and
C형간염항체=음성 and 혈소판>=130($10^3$/uL) and
알파태아성단백<=20(IU/L)

이다. [그림 2]은 SPAN에서 상자그림과 해당변수의 오즈비를 나타낸 것이다.

그림 2. SPAN에서 상자그림과 해당 오즈비



[표 9]는 앞에서 언급된 4가지 방법들에서 선택된 변수들을 정리한 것이다. 선택된 변수들을 살펴보면 HBsAg, AntiHCV, Family history가 모든 방법에서 선택되었고 drinking과 Alk.phos가 SPAN을 제외한 3가지 방법에서 선택되었다. 그밖에 platelet과 LYM, $\alpha$-FP가 선택되었다.

표 9. 간경변 발생 위험 인자

| Model | Risk factor | | | |
|---|---|---|---|---|
| Logistic regression | HBsAg<br>Alk.phos | AntiHCV<br>platelet | Family history<br>LYM | drinking |
| Polychotomous regression | HBsAg<br>Alk.phos | AntiHCV | Family history | drinking |
| QUEST | HBsAg<br>Alk.phos | AntiHCV | Family history | drinking |
| SPAN | HBsAg<br>$\alpha$-FP | AntiHCV | Family history | platelet |

## 4.6 간경변증 발생 위험군 분류 결과

[표 10]은 훈련용 자료를 가지고 4가지 모형에서 선택된 변수들을 가지고 간경변 발생에 대한 민감도, 특이도, 위음성율, 위양성율, 정확도를 구하여 정리한 것이다. 전체적인 정확도는 다항수준 회귀분석이 가장 높았고, QUEST와 로지스틱회귀분석이 그 뒤를 따랐다. SPAN의 정확도는 가장 낮았으나, 민감도가 다른 방법에 비해 높게 나왔고 위음성률도 다른 방법들의 절반정도의 수준이었다. [표 11]은 검증용 자료에서의 분석 결과로 훈련용 자료에서의 결과와 큰 차이를 보이지는 않았다. 여기에서도 SPAN의 민감도는 높으나 정확도가 약간 낮았다.

표 10. 간경변증 위험군 분류 결과(훈련용 자료)

| Model | Sensitivity | Specificity | False negative | False positive | Total accuracy |
|---|---|---|---|---|---|
| Logistic regression | 40.2 | 98.9 | 59.8 | 1.1 | 91.7 |
| Polychotomous regression | 54.6 | 97.3 | 45.4 | 2.7 | 92.1 |
| QUEST | 43.4 | 98.7 | 56.6 | 1.3 | 91.9 |
| SPAN | 78.5 | 89.0 | 21.5 | 11.0 | 87.7 |

표 11. 간경변증 위험군 분류 결과(검증용 자료)

| Model | Sensitivity | Specificity | False negative | False positive | Total accuracy |
|---|---|---|---|---|---|
| Logistic regression | 42.8 | 99.1 | 57.2 | 0.9 | 92.2 |
| Polychotomous regression | 55.6 | 97.3 | 44.4 | 2.7 | 92.2 |
| QUEST | 42.8 | 98.4 | 57.2 | 1.6 | 91.6 |
| SPAN | 81.6 | 89.1 | 18.4 | 10.9 | 88.2 |

## 4.7 공통변수 분석결과

로지스틱 회귀분석, 다항수준 회귀분석, QUEST, SPAN의 변수 선택 결과 HBsAg, AntiHCV, Family history가 공통적으로 선택되었다. 이 3가지 변수를 가지고 실시한 분류 결과를 보면 훈련용 자료에서 로지스틱 회귀분석과 다항수준 회귀분석에서 민감도가 36.7로 같게 나왔고, QUEST의 민감도는 29.5로 가장 낮고, SPAN의 민감도가 76.9로 나머지 분류 결과에 비해 2배 정도 높게 나타났다. 정확도는 SPAN의 경우 88.5로 로지스틱 회귀분석이나 다항수준 회귀분석의 91.0보다 낮게 나타났다. [표 12]와 [표 13]은 훈련용 자료와 검증용 자료에서 HBsAg, AntiHCV, Family history 3개의 변수를 가지고 자료를 분류한 결과를 정리한 것이다.

표 12. 4가지방법*에서 공통적으로 선택된 변수 분석결과(훈련용 자료)

| Model | Sensitivity | Specificity | False negative | False positive | Total accuracy |
|---|---|---|---|---|---|
| Logistic regression | 36.7 | 98.6 | 63.3 | 1.4 | 91.0 |
| Polychotomous regression | 36.7 | 98.6 | 63.3 | 1.4 | 91.0 |
| QUEST | 29.5 | 99.4 | 70.5 | 0.6 | 90.8 |
| SPAN | 76.9 | 90.1 | 23.1 | 9.9 | 88.5 |

* Logistic regression, Polychotomous regression, QUEST, SPAN

표 13. 4가지방법*에서 공통적으로 선택된 변수 분석결과(검증용 자료)

| Model | Sensitivity | Specificity | False negative | False positive | Total accuracy |
|---|---|---|---|---|---|
| Logistic regression | 40.4 | 98.6 | 59.6 | 1.4 | 91.5 |
| Polychotomous regression | 40.4 | 98.6 | 59.6 | 1.4 | 91.5 |
| QUEST | 32.8 | 99.3 | 67.2 | 0.7 | 91.2 |
| SPAN | 80.8 | 90.3 | 19.2 | 9.7 | 89.1 |

* Logistic regression, Polychotomous regression, QUEST, SPAN

다음은 로지스틱 회귀분석, 다항수준 회귀분석, QUEST에서 공통적으로 선택된 변수 HBsAg, AntiHCV, Family history, drinking, Alk.phos를 가지고 분류한 결과를 [표 14] 와 [표 15]에 정리하였다. 표를 보면 훈련용 자료의 민감도는 SPAN이 86.9로 가장 높았고, 특이도는 QUEST가 98.7로 가장 높았다. 정확도는 다항수준 회귀분석이 92.1로 높게 나타났다. 검증용 자료에서도 민감도는 SPAN이 89.2로 제일 높았고, 특이도는 QUEST가 98.4로 제일 높고, 정확도는 다항수준 회귀분석이 92.2로 높게 나타났다.

표 14. 3가지방법*에서 공통적으로 선택된 변수 분석결과(훈련용 자료)

| Model | Sensitivity | Specificity | False negative | False positive | Total accuracy |
|---|---|---|---|---|---|
| Logistic regression | 48.2 | 97.7 | 51.8 | 2.3 | 91.6 |
| Polychotomous regression | 54.6 | 97.3 | 45.4 | 2.7 | 92.1 |
| QUEST | 43.4 | 98.7 | 56.6 | 1.3 | 91.9 |
| SPAN | 86.9 | 72.6 | 13.1 | 27.4 | 74.3 |

* Logistic regression, Polychotomous regression, QUEST

표 15. 3가지방법*에서 공통적으로 선택된 변수 분석결과(검증용 자료)

| Model | Sensitivity | Specificity | False negative | False positive | Total accuracy |
|---|---|---|---|---|---|
| Logistic regression | 49.6 | 97.9 | 50.4 | 2.1 | 92.0 |
| Polychotomous regression | 55.6 | 97.3 | 44.4 | 2.7 | 92.2 |
| QUEST | 42.8 | 98.4 | 57.2 | 1.6 | 91.6 |
| SPAN | 89.2 | 71.7 | 10.8 | 28.3 | 73.9 |

* Logistic regression, Polychotomous regression, QUEST

다음은 SPAN에서 선택된 HBsAg, AntiHCV, Family history, platelet, LYM, $\alpha$-FP를 가지고 분류한 결과를 [표 16], [표 17]에 정리하였다. 훈련용 자료와 검증용 자료에서 SPAN의 민감도가 QUEST에 비해 3배 가까이 높았다. 반면 훈련용 자료에서 SPAN의 특이도는 89.0로 QUEST의 99.4보다 낮게 나타났다. SPAN의 정확도는 훈련용 자료에서 87.7, 검증용 자료에서 88.2로 나머지 세가지 방법들에 비해 약간 떨어졌다.

표 16. SPAN에서 선택된 변수를 대상으로 한 분석결과(훈련용 자료)

| Model | Sensitivity | Specificity | False negative | False positive | Total accuracy |
|---|---|---|---|---|---|
| Logistic regression | 42.2 | 98.3 | 57.8 | 1.7 | 91.4 |
| Polychotomous regression | 42.2 | 98.3 | 57.8 | 1.7 | 91.4 |
| QUEST | 29.5 | 99.4 | 70.5 | 0.6 | 90.8 |
| SPAN | 78.5 | 89.0 | 21.5 | 11.0 | 87.7 |

표 17. SPAN에서 선택된 변수를 대상으로 한 분석결과(검증용 자료)

| Model | Sensitivity | Specificity | False negative | False positive | Total accuracy |
|---|---|---|---|---|---|
| Logistic regression | 46.4 | 98.3 | 53.6 | 1.7 | 92.0 |
| Polychotomous regression | 45.2 | 98.4 | 54.8 | 1.6 | 91.9 |
| QUEST | 32.8 | 99.3 | 67.2 | 0.7 | 91.2 |
| SPAN | 81.6 | 89.1 | 18.4 | 10.9 | 88.2 |

마지막으로 로지스틱 회귀분석, 다항수준 회귀분석, QUEST, SPAN에서 뽑힌 HBsAg, AntiHCV, Family history, drinking, Alk.phos, platelet, LYM, $\alpha$-FP를 가지고 분석하였다. [표 18]과 [표 19]를 보면 훈련용 자료에서 SPAN의 민감도가 89.2로 가장 높았으나 정확도에서는 67.7로 다른 방법들에 비해 떨어졌다. 검증용 자료에서도 마찬가지고 SPAN의 민감도가 89.6으로 가장 높았고 정확도는 67.1로 다른 방법들에 비해 낮았다.

표 18. 선택된 모든 변수를 대상으로 한 분석결과(훈련용 자료)

| Model | Sensitivity | Specificity | False negative | False positive | Total accuracy |
|---|---|---|---|---|---|
| Logistic regression | 47.8 | 98.2 | 52.2 | 1.8 | 92.0 |
| Polychotomous regression | 54.6 | 97.3 | 45.4 | 2.7 | 92.1 |
| QUEST | 55.4 | 97.5 | 44.6 | 2.5 | 92.3 |
| SPAN | 89.2 | 64.7 | 10.8 | 35.3 | 67.7 |

표 19. 선택된 모든 변수를 대상으로 한 분석결과(검증용 자료)

| Model | Sensitivity | Specificity | False negative | False positive | Total accuracy |
|---|---|---|---|---|---|
| Logistic regression | 49.2 | 98.3 | 50.8 | 1.7 | 92.3 |
| Polychotomous regression | 55.6 | 97.3 | 44.4 | 2.7 | 92.2 |
| QUEST | 54.0 | 97.1 | 46.0 | 2.9 | 91.8 |
| SPAN | 89.6 | 63.9 | 10.4 | 36.1 | 67.1 |

전반적으로 SPAN의 민감도는 나머지 3가지 방법들에 비해 월등히 뛰어나나 선택된 변수가 늘어날수록 정확도가 떨어지는 것을 확인할 수 있었다.

# 제 5장 결론 및 고찰

통계학적으로 분류를 하는 방법인 SPAN에 대한 유용성 평가를 위해, 분류를 위한 분석에 쓰이는 로지스틱 회귀분석과 다항수준 회귀분석, 나무 모형을 이용한 QUEST를 이용하여 간경변 발생 위험군 분류 결과를 비교해 보았다.

본 논문에서는 1994년부터 2005년 사이에 건강 검진을 받은 자 중에서 병원에 내원하여 간경변 발생 여부에 대한 진단을 받은 4,093명의 자료를 이용하였다.

분석 결과 SPAN의 민감도가 다른 방법들에 비해 약 2배 정도 높은 것으로 나타났다. 민감도는 실제 간경변인 자를 간경변 위험군으로 분류한 정도를 나타내는 것으로 임상 의학에서 간경변 위험군 분류시 유용할 것으로 생각된다. 전체적인 정확도가 다른 방법들에 비해 4~5% 떨어지지만 임상 의학적으로 분류의 결과에 대한 오분류 비용을 고려한다면 민감도가 높은 SPAN은 다른 로지스틱 회귀분석이나 QUEST의 결과보다 유용할 것이다. SPAN은 민감도가 높은 것과 함께 해석이 쉽다는 장점을 가지고 있다. 분류결과의 해석력은 중요하다. 특히 의학적인 진단과 예측의 규칙이 복잡하면 사용이 불가하다. 그리고 SPAN은 종속변수에 대한 양의 속성만의 결합으로 이루어지기 때문에 애매한 해석이 나오지 않는다는 장점이 있다. QUEST의 분류결과를 보면 B형간염 바이러스가 양성이고, 가족력이 없고, 알칼리성 포스타파제가 115IU/L 이하이면 간경변증 비위험군으로 분류되었다. 하지만 B형간염 바이러스는 간경변증에 가장 유의미한 변수로 알려져 있다. 이런 분류의 결과는 이해하기 힘들게 만든다. SPAN의 알고리즘은 이런 애매하고 이해하기 힘든 결과를 사전에 예방한다. 게다가 SPAN은 독립변수와 종속변수 모두 이분형으로만 설정하여 분석을 실시한다. 이는 임상 의학에서 특정 검사 후 연속적인 값을 정상/비정상으로 나누는 성격을 잘 반영할 것이다.

SPAN의 통계학적 유용성 뿐만 아니라 SPAN에서 제시되는 상자그림도 유용하다고 보여진다. 임상의학분야에서 위험군을 시각적으로 보여주는 방법은 잘 발전하지 못하였고, 보통은 표를 이용하여 나타낸다. SPAN에서 사각형의 크기를 이용하여 위험군을 나타내는 방법은 매우 유용한 접근이다. 이렇게 보여주는 방법은 위험 요인들

의 관계를 보여주는 점과 특별히 더 위험한 결합을 나타내는데 유용할 수 있다.

SPAN의 장점이 있지만 몇 가지 제한점도 있다. 첫 번째로 종속변수의 범주가 3개 이상이면서 순서형인 경우에 적용할 수 없다. 두 번째로 로지스틱회귀분석이나 다항 수준회귀분석에서는 사후확률을 이용한 분리점을 통해 민감도, 특이도, 정확도의 최적의 분리점을 제시해 주는데, SPAN에서는 최선의 분리기준을 한번에 얻기 어렵다. 연구자가 부울결합의 크기를 임의로 변화해 가면서 찾아야만 한다. 세 번째로 속성의 분리점을 연구자가 정해주어야 한다는 점이다. QUEST는 독립변수의 가능한 분리기준을 모두 탐색한 후 최상의 분리점을 제시하는데, SPAN에서도 이런 알고리즘의 적용이 필요하다고 본다.

# 참고 문헌


성웅현. *응용 로지스틱 회귀분석*. 2001

정혜원. 통계적 기법과 데이터마이닝 기법을 이용한 이동통신 VAS 가망고객 scoring 모형 비교 연구. 2004

하정윤. MDR을 이용한 간경변증 발생 고위험군 분류. 2007

Austin, P. C. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Statistics in Medicine*, 2007:26;2937-57.

Chaudhuri, P., Lo, W. D., Loh, W. Y., Yang, C. C. Generalized regression trees. *Statistica Sinica*, 1955:5;641-66.

Kooperberg, C., Bose, S., Stone, C. J. Polychotomous Regression. *Journal of the American Statistical Association*, 1997:92;117-27.

Loh, W. Y., Shih, Y. S. Split selection methods for classification trees. *Statistica Sinica*, 1997:7;815-40.

Lim, T. S., Loh, W. Y., Shih, Y. S. A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. *Machine Learning*, 2000:40;203-28.

Marshall, R. J. Partitioning methods for classification and decision making in medicine. *Statistics in Medicine*, 1986:5;517-26.

Marshall, R. J. A program to implement a search method for identification of clinical subgroups. *Statistics in Medicine*, 1995:14;2645-59.

Marshall, R. J. Determining and visualising at-risk groups in case-control data. *Journal of Epidemiology and Biostatistics*, 2001:6;343-8.

Marshall, R. J. The use of classification and regression trees in clinical epidemiology. *Journal of Clinical Epidemiology*, 2001:54;603-9.

Marshall, R. J. Comparison of misclassication rates of search partition analysis and other classication methods. *Statistics in Medicine*, 2006:25;3787-97.

Picard, R. R., Berk, K. N. Data Splitting. *The American Statistician*, 1990: 44;140-7.

Su, X. Tree-based medel checking for logistic regression. *Statistics in Medicine*, 2007:26;2154-69.

Zhang, H., Holford, T., Bracken, M. B. A tree-based method of analysis for prospective studies. *Statistics in Medicine*, 1996:15;37-49.

# ABSTRACT

## Assessment of utility of SPAN for classification of risk group for the development of liver cirrhosis

You, Young Ae

Dept. of Biostatistics and Computing

The Graduate School

Yonsei University

The statistical predictive methods were used to find the risk factors related with disease and to generate predictive probabilities of those diseases. Logistic regression is the most commonly used method for predicting the probability of diseases in the medical fields. Also, data-driven methods, such as classification and regression trees(CART) have been used to identify subjects at increased risk of diseases. However, both of regression and tree models have their specific limitations in spite of their advantages.

Recently, an alternative approach called by search partition analysis(SPAN) is suggested, which is based on direct non-hierarchical search algorithm to identify subgroups at risk. SPAN searches subgroups among different Boolean combinations of risk factors. It was known that SPAN had an advantage that its decision rules are usually more interpretable than those of other methods, especially in medical problems that interpretability of decision rules is very important.

In this thesis, SPAN was compared against the performance of the other 3

methods; logistic regression, polychotomous regression and quick unbiased efficient statistical trees. We applied these methods to the real clinical data composed of 4,093 individuals who received the screening test in first and then visited Yonsei University Medical Center for check-up liver cirrhosis from May 1994 to September 2005. The performance of SPAN and that of any other methods were compared and the measures of performance were sensitivity, specificity, and accuracy. In the results using SPAN, the findings identified by the risk factors for liver cirrhosis were HbsAg, AntiHCV, Family history, platelet and $\alpha$-FP. And we found that the sensitivity using SPAN were much higher than those of other methods in various data sets.

In conclusion, as long as it works, the performance of SPAN should make sense in the context of medical diagnosis and prognosis.

---

Key words : SPAN, Logistic regression, polychotomous regression, QUEST, Sensitivity, Specificity, Accuracy