PRIM을 이용한
고혈압 질환 발생의 유전적 위험군 분류

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

안 진 희

PRIM을 이용한
고혈압 질환 발생의 유전적 위험군 분류

지도 장 양 수 교수

이 논문을 석사 학위논문으로 제출함

2007년 12월 일

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

안 진 희

안진희의 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

PRIM을 이용한
고혈압 질환 발생의 유전적 위험군 분류

고혈압은 복합성 질환(complex disease)으로 다양한 유전자 변이 및 환경적 요인이 복합적으로 질병 발생과 기전에 영향을 미치는 것으로 알려져 있다. 따라서 고혈압 발생에 대한 위험 인자를 찾기 위한 연구 수행을 위해서는 다요인적 위험요인을 평가하기 위한 분석 방법을 필요로 한다.

본 연구에서는 복합적 위험인자를 고려한 고혈압 발생에 대한 유전적 위험군을 분류하고자 했는데, 이를 위해 PRIM(patient rule induction method)을 적용하였다. 이 방법은 비유전적 요인과 유전요인, 환경요인 등 다중 요인의 영향에 따른 질병 발생 위험군을 구체적으로 분류 할 수 있다는 장점을 가지고 있다. 실제자료로 심혈관계유전체 센터에서 조사된 고혈압 환자군 560명, 정상대조군 378명, 총 938명을 대상으로 하여, 첫 번째 단계에서는 전통적으로 고혈압 질환에 영향을 미치는 나이, 성별, 음주여부, 운동여부, 당뇨병여부의 위험인자로 분석을 수행하였고, 두 번째 단계에서는 혈중지질농도, 체질량 지수를 위험인자로, 마지막 단계에서는 최종적인 유전적 위험군 분류를 위해 단일염기다형성을 위험인자로 하여 분석을 수행하였다. 그리고 이 결과를 기존의 위험군 예측 방법으로 흔히 사용되는 로지스틱 회귀분석을 적용한 결과와 비교하였다. PRIM을 적용한 결과, 기존의 분류 방법과는 달리 위험요인들을 차례로 추가시켰을 때의 위험도와 그에 따른 위험군을 분류할 수 있었으며, 또한 현실적으로 자료를 분석하는 데 있어서 임상적으로 의미 있는 해석이 가능하다는 점을 확인 할 수 있었다.

핵심 되는 말 : 고혈압, patient rule induction method, 로지스틱 회귀분석

제 1장 서론

1.1 연구 배경

대부분의 심혈관계질환은 복합성 질환(complex disease)으로 다양한 유전자 변이가 복합적으로 질병 발생과 기전에 영향을 미치는 것으로 알려져 있다. 따라서 다요인적인 위험 요인을 평가하기 위한 방법을 필요로 한다. 또한 복합성 질환 발생과 관련하여 알려진 후보 유전자의 다형성이 질병 발생과 관련된 독립적인 위험인자는 아닐지라도 기존의 알려진 위험요인에 의해 위험도를 높이는 것으로 보고되어져 있다.

명확한 유전적 원인에 대해서는 질병의 과거력이나 확실한 가족력에 의해 확인되어 질 수 있지만, 이러한 일반적인 모집단에 대한 질병과 관련된 유전자와 다른 위험 요소와의 연관성에 대해서는 충분히 설명될 수 없다. 질병의 유전적 경향은 하나의 단일 유전자의 영향이기 보다는 환경적 요인이 고려되어져 질병의 위험도를 높이기 때문이다.

따라서 질환 발생의 예측과 예방을 위해 유전적 위험요인과 환경적 위험 요인을 구체적으로 분류할 수 있는 접근 방법이 요구된다.

1.2 연구목적 및 내용

지금까지 human genetics 분야에서 질병 발생에 위험을 미치는 다인자성 요인을 토대로 특별한 질병 발생에 대한 위험군을 분류(classification) 하는 여러 가지

통계학적 방법들이 제안되었는데, 이러한 방법을 임상자료 분석에 사용하여 그 효과를 평가해 보고자 한다.

본 연구에서는 질병 발생에 대한 위험군을 분류하기 위해 원 자료(raw data)에 reference 범주를 생성하여 분류하는 방법인 PRIM(patient rule induction method) 방법을 적용하고자 하는데, 이 방법은 비유전적요인과 유전요인, 환경요인 등 다중요인의 영향에 따른 질병 발생의 차이를 확인 할 수 있고, noise 변수가 포함 된 경우에도 분류가 잘 된다는 특징으로 인해 질병의 위험군을 좀 더 구체적으로 분류 할 수 있다는 장점을 가지고 있다. 또한 기존의 방법 중 가장 대표적인 방법으로 모수적 선형 회귀분석 방법인 로지스틱 회귀분석을 PRIM의 방법과 비교하여 보고자 한다.

실제 비교, 평가를 위해 본 연구에서는 심혈관계질환 유전체연구센터에서 조사된 자료로 고혈압 질환 진단을 받은 환자군과 정상 대조군의 유전적 정보 및 환경적 정보를 이용하고자 한다.

1.3 논문 구성

제 1장 서론에서는 연구의 배경에 대해 소개하고 연구 목적 및 내용에 대해 제시한다. 2장에서는 기존의 통계적 방법 중에 로지스틱 회귀분석의 이론적 배경과 방법에 대해 설명한다. 3장에서는 patient rule induction method의 이론적 내용을 소개하였다. 4장에서는 분석에 사용된 심혈관계질환 유전체연구센터에서 조사된 자료를 설명하고, 2장에서 소개한 방법으로 비교 분석하였으며, 5장에서는 결론 및 고찰에 대해서 논의하고, 앞으로 진행되어야 할 연구에 대한 제안점을 서술하였다.

제 2장 분류 방법

2.1 로지스틱 회귀분석(Logistic Regression)

2.1.1 이론적 배경

반응 변수의 값이 연속적이지 않고 “성공”, “실패” 나 “생존”, “사망” 등의 두 개의 범주이거나 또는 그 이상의 범주로 나누어져 있는 경우에 단순회귀모형을 적용하면 반응변수를 연속형으로 간주하기 때문에 기존의 가정을 만족하지 못한다. 이 때 성공확률을 추정하고, 그 값에 유의한 영향을 미치는 설명변수가 무엇인가를 알아보기 위해 사용되는 분석이 로지스틱 회귀분석(logistic regression)이다.

2.1.2 로지스틱 회귀모형

여러 설명변수로부터 두 범주만을 가지는 종속변수를 예측하는데 사용되는 로지스틱 회귀분석은 모형구조에 의해 연관성 및 교호 작용 유형을 설명할 수 있으며 모수의 추론을 통해서 반응 값에 대한 독립변수의 영향력을 평가할 수 있다.

p개의 설명변수 $x_{1,}x_{2,}\cdots,x_p$ 에 대하여 로지스틱 회귀모형은 다음과 같다.

$$\log\frac{p(y=1|x_{1,}\cdots,x_p)}{1-p(y=1|x_{1,}\cdots,x_p)}=\beta_0+\beta_1x_1+\cdots+\beta_px_p$$

위의 식은 모수 β_i에 대해서 일차 선형관계를 가지며, 관찰치 y가 1로 분류될 사후확률인 $P(Y=1|x_{1,}x_{2,}\cdots x_p)$에 대해서 정리하면 다음과 같다.

$$P(Y=1|x_{1,}x_{2,}\cdots x_p)=\frac{\exp(\beta_0+\beta_1x_1+\cdots+\beta_px_p)}{1+\exp(\beta_0+\beta_1x_1+\cdots+\beta_px_p)}$$

이것을 로지스틱 반응함수(logistic response function)라고 한다. 이렇게 얻어진 각각의 사후확률은 0과 1사이의 값을 가지게 되고, 해당 객체를 분류하기 위해 적절한 절단값(cutoff value)을 정하여 이 값을 기준으로 분류한다.

로지스틱 회귀분석은 분류기법으로 가장 널리 사용되고 있는 방법으로 선형성(linearity)을 가정함으로써 해석이 용이하며 회귀계수나 오즈비(odds ratio) 같은 분석의 결과는 많은 유용한 정보를 제공한다. 그러나 각 설명변수의 영향이 다른 설명변수에 종속되어 있지 않다고 가장함으로써 일부 변수들 간의 교호작용(interaction)을 발견할 수가 없으며, 변수들 간의 관계가 복잡한 비선형성의 자료인 경우 예측의 측면에서 한계를 가진다.

제 3장 Patient Rule Induction Method

3.1 이론적 배경

Friedman과 Fisher(1997)는 규칙에 의한 군집화(clustering)와 목적함수(object function) 값의 최적화를 동시에 실시하면서 오차를 최소화시킨 PRIM(Patient Rule Induction Method, PRIM)이라는 알고리즘을 개발하였다. 이 방법의 특징은 다차원의 자료를 공간상의 상자(box) 형태로 이해하고 그 상자를 조금씩 잘라나갈 때 남아있는 상자 안의 자료값에 대한 목적 함수값을 계산하여, 그 함수값이 최대화 또는 최소화되는 시점의 상자에 대한 규칙을 발견하는 것이다. 이 방법은 다중요인의 영향에 따른 차이를 확인 할 수 있고, noise 변수가 포함 된 경우에도 분류가 잘 된다는 특징으로 인해 기존의 방법들 보다 좀 더 구체적으로 분류 할 수 있다는 장점을 가지고 있다.

3.2 알고리즘

많은 자료 분석이나 분류의 목적은 정확한 예측이다. 분석 대상인 자료의 구조는 종속변수 y에 대하여 동시에 측정된 입력변수들 $\boldsymbol{x}=(x_1, x_2, \cdots, x_n)$의 반복으로 구성되어 있다. 다시 말하면 자료 분석의 목적은 자료

$$\{y_i, \boldsymbol{x_i}\}_1^N$$

를 이용하여 입력 x에 대하여 실제 값 y와 비슷한 값을 계산하는 목적함수(target function) $f(x)$를 추정하는 것이다. 이와 같은 함수추정 문제는 목적함수 $f(x)$의 최대 또는 최소값을 찾는 문제에 응용되기도 한다. 본 연구에서는 전체 평균보다 훨씬 큰 평균 영역(region)을 찾는 것이다.

설명변수 x_j가 취할 수 있는 모든 가능한 값들의 집합을 S_j라고 한고 집합 S를 전체 설명변수들의 가능한 집합이라고 하면

$$S = S_1 \times S_2 \cdots \times S_n$$

로 표현된다. 목적은 집합 S의 다음 조건을 만족하는 $R(\subset S)$을 찾는 것이다.

$$\begin{aligned} \overline{f_R} &= average_{x \in R} \bar{f}(\boldsymbol{x}) \\ &= \int_{\boldsymbol{x} \in R} f(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} / \int_{\boldsymbol{x} \in R} p(\boldsymbol{x}) d\boldsymbol{x} \gg \bar{f} \end{aligned}$$

여기서 $\bar{f}$는 존체 영역에서 목적함수의 평균이다. 부분집합 R의 크기는 다음과 같이 정의 할 수 있다.

$$\beta_R = \int_{\boldsymbol{x} \in R} p(\boldsymbol{x}) d\boldsymbol{x}$$

여기서 $p(\boldsymbol{x})$는 설명변수들 $\boldsymbol{x}$의 확률밀도 함수이다. 그러므로 β_R, 즉 집합 R의 크기는 해당영역의 확률을 의미한다. 결론적으로 함수 f_R가 최대가 되는 영역 R을 찾아야 하는데 R의 크기가 작으면 작을수록 f_R의 값은 커지고 반대로 관심영역의 크기 β_R이 크면 함수값이 작아지는 편향분산의 균형문제(bias-variance trade-off)가 일어난다. 실제 사용하는 β_R과 $\overline{f_R}$의 추정치는 다음과 같다.

$$\widehat{B_R}= \frac{1}{N}\sum_{x_i \in R} I(x_i \in R),\ \overline{Y_R}= \frac{1}{N \bullet \widehat{B_R}}\sum_{x_i \in R} y_i$$

B_k들의 합집합으로 R을 표현 가능하다고 하면

$$R= \bigcup_{k=1}^{K} B_k$$

로 나타낼 수 있다.

집합 s_{jk}를 변수 x_j의 집합으로써, 변수 x_j의 모든 가능한 값들의 집합인 S_j의 부분집합이라 하면 결국 B_k는 공간상에 상자(box)의 모형을 지닌 집합으로

$$B_k = s_{1k} \times s_{2k} \times \cdots \times s_{nk} = \bigcup_{j=1}^{n} (x_j \in s_{jk})$$

로 표현 할 수 있다. 규칙 R을 구성하는 $\{B_k\}_1^K$에서, 첫 상자인 B_1은 전체 설명변수영역에서 도출된 것이고 두 번째 상자 B_2는 B_1를 제거한 후 구해진 부분 영역이다. 상자 B_K는 이전의 $K-1$개의 상자가 제거된 후 도출된 집합이다. 찾아야 할 영역 R은 평균값이 어떤 기준값 $\overline{y_0}$보다 큰 상자들의 합집합이거나 혹은 편향분산의 균형으로 인해 영역 R은 지정된 영역의 크기 β_t에 대하여 최대의 평균값 $\overline{y_R}$을 갖는 상자들의 집합으로 생각할 수 있다.

$$\beta_R = \sum_{k=1}^{K} \beta_k \simeq \beta_t$$

전체영역에서 보다 훨씬 큰 평균을 갖는 영역 R은 해석이 가능한 규칙들의 집

합인 $\bigcup_{k=1}^{K} B_k$로 표현된다. 다음으로 규칙들의 집합인 $\{B_k\}_{k=1}^{K}$을 도출하는 방법은 편향분산의 균형문제가 있다. 이를 해결하기 위해 PRIM의 알고리즘은 기본적으로 다음의 top-down peeling과 buttom-up pasting 두 단계[그림 1]로 나누어진다. 여기서 N은 관측자료 집합의 수이고 θ는 누적 발생률 사건이다.

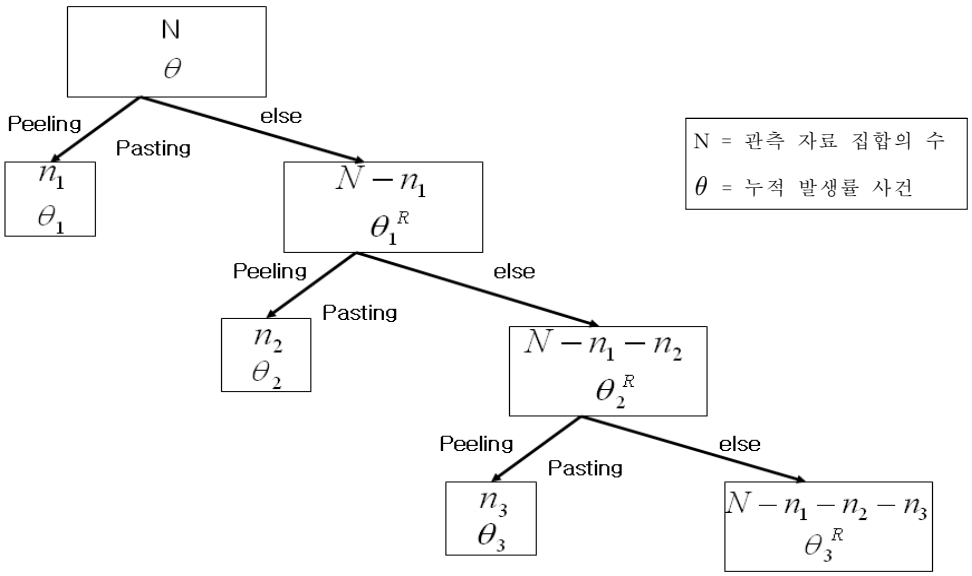

그림 1. PRIM 알고리즘

3.2.1 Top-down peeling

Top-down peeling은 전체 분리의 영역을 상자 B라고 하면 반복적인 과정을 통해 B에 속하는 상자 b를 제거(peeling) 하면서 남아있는 상자 $B-b$ 안의 평균값이 최대가 되도록 하는 방법이다.

자세한 알고리즘은 다음과 같다

1) 전체 자료에서 subbox b^*가 제거되었을 경우 가장 큰 결과 평균을 산출하게 하는 subbox b^*를 찾는다.

$$b^* = \arg max_{b \in c(b)} ave[y_i | x_i \in B-b]$$

2) $C(b)$는 각각의 설명변수에서 정의된 상자들을 모두 합한 것으로 해석 가능성에 의해 제한을 받는다. subbox b^*는 각각의 설명변수 x_j에 의해 정의된다. 설명변수 x_j가 실수인 경우 subbox b^*는 b_{j-}와 b_{j+} 중 하나이다. b_{j-}는 현재 상자에서 j 번째 설명 변수의 하한 $\{x | x_{j(} < x_j(\alpha)\})$이고, b_{j+}는 현재 상자에서 j번째 설명 변수의 상한 $\{x | x_j > x_j(1-\alpha)\}$이다. $x_j(\alpha)$는 현재 상자 안에서 자료 x_j값의 α분위수이다. 일반적으로 $\alpha < 0.1$이다

설명변수 x_j가 범주형인 경우 subbox b^*는 j번째 변수의 m번째 범주의 형태이다.

$$b_{jm} = \{x | x_j = s_{jm}\}; s_{jm} \in S_j$$

3) 위에서 찾은 적절한 subbox b^*를 제거하여 상자를 갱신한다.

$$B = B - b^*$$

4) 상자에서 관측 값의 최소의 수 β_0가 남겨질 때까지 단계1과 단계2를 반복한다.

$$stop : \beta_B = \frac{1}{N}\sum_{i=1}^{N} 1(x_i \in B) \leq \beta_0$$

3.2.2 Bottom-up pasting

top-down peeling의 목표는 결과 평균의 큰 설명 변수의 부분 영역을 찾는 것이다. 상자의 경계선은 top-down peeling절차의 많은 단계에서 선택된 부분상자를 정의하는 변수들의 특정한 값이다. 맨 마지막 상자를 제외한 각 과정에서 선택된 상자의 경계선은 다른 변수에서 경계선을 더 정제하는 이후의 주변 제거에 관한 지식 없이 바로 전 단계에 의해서 결정된다. 이것은 최종적인 상자가 경계선을 다시 조정함으로써 개선될 수 있는 가능성이 있다는 것으로 bottom-up peeling 과정을 하는 이유가 된다.

알고리즘은 기본적으로 top-down peeling의 역으로 다음과 같다.

1) peeling solution B를 가지고 시작한다.

2) 적절한 subbox b^*를 본래의 상자 B에 더하여 상자를 갱신한다.

$$B = B \bigcup b^*$$

b^*는 새로운 커진 상자에서 결과 변수의 평균을 최대화시키는 적절한 영역으로 top-down peeling에서 정의되는 것과 같다. 설명변수 x_j가 실수인 경우 적당한 subbox는 x_j의 B의 상한과 하한을 확장하는 구간으로 나타난다. 이 구간의 넓이는 peeling fraction α와 현재의 상자 안에 있는 관측 값의 수 N_B의 곱인 αN_B 개의 관측치를 포함하는 것으로 선택된다. 설명변수 x_j가 범주형인 경우 현재의 상자에서 나타나지 않은 s_{jm}으로 더해지기 적당한 subbox를 정의한다.

3) 다음 subbox b^*가 더해지는 것이 결과변수의 평균 $\bar{y}_{B+b^*}$를 감소시킬 때까지 2)를 반복한다.

Bottom-up pasting은 peeling solution를 개선할 수 있는 기회가 되기는 하지만 극적인 영향을 가지지는 않는다.

제 4장 실제자료를 이용한 고혈압 질환 발생의 유전적 분류

4.1 심혈관계질환 유전체연구센터 자료

분석에 사용한 자료는 연세대학교 심혈관계질환 유전체 연구센터에서 실시하는 검진에 동의한 대상자들의 임상적 자료를 이용하였다. 자료는 고혈압, 관상동맥질환 등의 질병이 없는 대조군과 고혈압으로 진단 받은 고혈압 환자를 환자군으로 이루었다.

심혈관계질환 유전체 연구센터 자료에는 나이, 성별, 음주유무, 운동여부, 당뇨병 진단여부, 총 콜레스테롤(Total cholesterol, Tchol), 고밀도지단백콜레스테롤(High density lipoproteins, HDL), 저밀도지단백콜레스테롤(Low density lipoproteins, LDL) 중성지방(Triglyceride, Tg), 체질량 지수(Body Mass Index, BMI)등의 전통적으로 고혈압과 관련 있는 변수들과 7개의 염색체에 위치한 17개의 단일 염기 다형성(Single nucleotide polymorphism, SNP)들에 대해 조사하였다.

4.2 분석 대상 변수 선택

4214명의 심혈관유전체 데이터에서 고혈압으로 진단으로 받은 환자군 604명과 대조군 419명 총 1023명을 분석 대상으로 선정하였다. 단일 염기 다형성들과 위험요인들에 대한 연관성을 평가하기 위하여 위험요인 변수와 분석 대상 유전자를 선택한다. 기본적인 변수인 나이, 성별, BMI, 음주유무, 운동여부, 당뇨병 진단여부

는 전통적인 고혈압 위험 요인이므로 모두 분석 대상 변수에 포함한다. 그 외에 심혈관계 질환의 위험요인으로 알려진, Tchol, HDL, LDL, Tg를 고혈압 발생군과 비발생군간의 유의한 차이가 있는지 일변량 분석하여[표 1] 유의한 차이를 보이는 지질 위험 요소 HDL과 Tg를 선택하였다.

표 2 Lipids factors의 일변량 분석

Covariate	Without HT	With HT	p-value
Tchol	206.05±37.031	204.81±40.427	0.6171
HDL	49.10±12.193	44.33±11.729	0.0000*
LDL	131.62±33.307	130.03±37.061	0.4822
Tg	126.75±73.559	152.24±74.462	0.0000*

* p-value < 0.05

4.3 분석 대상 유전자 선택

고혈압 위험 요인과 관련 있는 유전자를 선택하기 위하여 이용 가능한 7개의 유전자에 위치한 17개의 단일 염기 다형성들에 대해 모두 조사하였다. 이들 SNP에 대한 정보는 다음의 [표 2]에 정리하였다.

표 3 SNP 변수 이름 목록

Alias	SNP	Gene name	location
ABCA1	R219K(G/A)	ATP-binding cassette, sub-family A(ABC1), member 1	9q31.1
ACE_1	A-240T(A2400T)	angiotensin I coverting enzyme (perptidy-dipeptidase A) 1	17q23
ACE_2	C-93T(C2547T)	angiotensin I I coverting enzyme (perptidy-dipeptidase A) 2	17q23
ACE_6	14094(I/D)	angiotensin I I coverting enzyme (perptidy-dipeptidase A) 6	17q23
ACE_7	G14480C	angiotensin I I coverting enzyme (perptidy-dipeptidase A) 7	17q23
ACE_8	T849S(A14519G)	angiotensin I I coverting enzyme (perptidy-dipeptidase A) 8	17q23
ACE_10	A22982G	angiotensin I I coverting enzyme (perptidy-dipeptidase A) 10	17q23
AGT_2	G-217A	angiotensinogen	1q42-q43
AGT_3	A-20C	angiotensinogen	1q42-q43
AGT_4	G-6A	angiotensinogen	1q42-q43
AGT_5	M235T(T/C)	angiotensinogen	1q42-q43
APOA5_2	T-1331C	apolipoprotein A-V	11q23
CETP_2	C-629A	cholesteryl I I ester transfer protein, plasma	16q21
CETP_3	TAQ1B(G/A)	cholesteryl I I ester transfer protein, plasma	16q21

표 2. SNP 변수 이름 목록(계속)

Alias	SNP	Gene name	location
CETP_5	I405V(A/G)	cholesteryl ester transfer protein, plasma	16q21
LDLR	N591N(C/T)	low density lipoprotein receptor	19p13.3
LIPC_4	V95M(G/A)	lipase, hepatic	15q21-q23

이들 SNP를 일변량 분석 하여 유의한 차이가 있는 SNP AGT_3, CETP_2, LDLR 세 가지 를[표 3] 유의한 확률 순서에 의해 선택하였다.

표 3. Genetic risk factors의 일변량 분석

Alias	Without HT	With HT	p-value
ABCA1			
AA	87	130	0.9389
GA	222	313	
GG	114	163	
ACE_1			
AA	132	211	0.4811
AT	204	277	
TT	87	118	
ACE_2			
CC	87	115	0.4121
CT	184	269	
TT	127	212	
ACE_6			
DD	73	101	0.8531
ID	205	287	
II	145	218	
ACE_7			
CC	145	213	0.904
CG	203	288	
GG	74	100	
ACE_8			
AA	147	217	0.8648
AG	205	281	
GG	70	101	

표 3. Genetic risk factors의 일변량 분석(계속)

Alias	Without HT	With HT	p-value
ACE_10			
AA	132	196	0.9241
GA	205	293	
GG	81	112	
AGT_2			
AA	8	16	0.5343
GA	130	199	
GG	285	391	
AGT_3			
AA	273	421	0.0093*
CA	128	175	
CC	19	9	
AGT_4			
AA	274	413	0.5197
GA	130	165	
GG	14	22	
AGT_5			
CC	279	416	0.613
CT	130	169	
TT	14	21	
APOA5_2			
AA	199	306	0.5517
GA	193	258	
GG	31	42	
CETP_2			
AA	86	173	0.0166*
CA	207	298	
CC	92	108	
CETP_3			
AA	60	86	0.5352
GA	202	309	
GG	161	211	
CETP_5			
AA	118	158	0.6262
GA	213	309	
GG	85	136	

표 3. Genetic risk factors의 일변량 분석(계속)

Alias	Without HT	With HT	p-value
LDLR			
CC	314	471	0.3426
CT	103	130	
TT	6	5	
LIPC_4			
AA	43	64	0.7762
GA	179	243	
GG	201	299	

* p-value < 0.05

앞서 선택한 전통적인 고혈압 위험 요인인 나이, 성별, BMI, 음주유무, 운동여부, 당뇨병 진단여부와 일변량 분석하여 선택한 지질요인 HDL, Tg 그리고 AGT_3, CETP_2, LDLR 세 가지의 SNP의 정보가 모두 있는 고혈압 환자군 560명 정상대조군 378명 총 938명을 최종 분석 대상으로 선정하였다.

4.4 변수의 구성

임상적 의미가 있는 PRIM 분석을 위하여 독립변수들이 범주형 자료로 이루어져야 한다. 심혈관계질환 유전체센터 자료의 나이의 평균이 정상군 46.80±10.28, 고혈압 환자군 55.51±10.84 이므로 45세 미만과 45세 이상으로 이분형 하였다. HDL값은 40mg/dl 미만과 이상으로 이분형 하였다(Cleeman et al, 2001). 또한 Tg 값도 150mg/dl 미만과 이상으로 이분형 하였다(Cleeman et al, 2001). BMI 값은 23kg/m^2 이하와 초과로 과체중 기준으로 이분형 하였다(대한비만학회, 2005). 이분형된 자료는 다음 [표 4]와 같다.

표 4. 변수들의 범주화 및 빈도

Covariate	Without HT (n=378)	With HT (n=560)
Traditional risk factor		
Age		
< 45	143(0.38)	91(0.16)
≥ 45	235(0.62)	469(0.84)
Sex		
Male	173(0.46)	242(0.43)
Female	205(0.54)	318(0.57)
Drinking		
No	145(0.38)	272(0.49)
Yes	233(0.62)	288(0.51)
Excercise		
No	251(0.66)	267(0.47)
Yes	127(0.34)	298(0.53)
Diabetes mellitus		
No	376(0.99)	542(0.97)
Yes	2(0.01)	18(0.03)
Lipids and BMI		
HDL	84(0.22)	209(0.37)
< 40	294(0.78)	351(0.63)
≥ 40		
Triglycerides		
< 150	270(0.71)	326(0.58)
≥ 150	108(0.29)	234(0.42)
BMI		
≤ 23	202(0.53)	158(0.28)
> 23	176(0.47)	402(0.72)
Genetic risk factors		
AGT_3		
AA	254(0.67)	390(0.70)
AC	106(0.28)	162(0.29)
CC	?18(0.05)	8(0.01)
CETP_2		
AA	85(0.22)	170(0.30)
AC	203(0.54)	284(0.51)
CC	90(0.24)	106(0.19)

표 4. 변수들의 범주화 및 빈도(계속)

Covariate	Without HT (n=378)	With HT (n=560)
LDLR		
CC	280(0.74)	436(0.78)
CT	4(0.25)	119(0.21)
TT	4(0.01)	5(0.01)

4.5 분석 결과

4.5.1 Patient rule induction method

PRIM방법을 이용하여 11개의 변수를 3단계로 적용하여 살펴본 결과 각각의 단계에서 위험 효과를 더하여 마지막 3단계에서 가장 높은 위험률을 보인다.

첫 번째 단계에서는 전통적으로 알려져 있는 고혈압의 위험 요인을 고려하여 4개의 분할을 만들었다. 두 번째 단계에서는 첫 번째 단계에서 분류한 위험 집단을 지질요인과 BMI 요인의 위험도를 가중시켜 각각을 2개 또는 3개의 분할을 만들었다. 세 번째 단계에서는 고혈압의 유전적 위험군을 분류하기 위하여 두 번째 단계에서 나눈 집단에 유전적 위험 요인을 가중시켜 고혈압의 고위험군을 보고자 했다. 다음의 [표 5]은 위에서 말한 3단계 PRIM 적용의 결과이며 [그림 2]는 결과를 그림으로 표현한 것이다.

표 5. 3단계 PRIM으로부터의 통계적 유의한 분할

Partition label	Traditional HT risk factor terms & Risk estimation(%)	Partition label	Lipids and BMI HT risk factor terms & Risk estimation(%)	Partition label	Genetic HT risk factor terms & Risk estimation(%)
P_1	Age≥45 & Male (66.09)	P_{11}	BMI>23 & HDL<40 (84.00)	P_{111}	CETP_2=AA (96.30)
				P_{11R}	not P_{111} (79.73)
		P_{1R}	not P_{11} (56.61)	P_{1R1}	CETP_2=AA (75.00)
				P_{1RR}	not P_{1R1} (48.87)
P_2	Age≥45 & Female & Drink & no-DM (60.34)	P_{21}	Tg>150 (70.45)	P_{211}	LDLR=CC (70.59)
				P_{21R}	not P_{211} (70.00)
		P_{2R}	not P_{21} (54.17)	P_{2R1}	AGT_3=AA (56.56)
				P_{2RR}	not P_{2R1} (51.85)
P_3	Age≥45 & Female & no-Drink & no-DM (69.57)	P_{31}	BMI>23 & HDL<40 (90.91)	P_{311}	AGT_3=AA& LDLR=CC (93.94)
				P_{31R}	not P_{311} (85.71)
		P_{32}	HDL≥40 (64.38)	P_{321}	LDLR=CC (66.67)
				P_{32R}	not P_{321} (57.89)
		P_{3R}	not P_{31} (68.00)	P_{3R1}	AGT_3=AA (68.75)
				P_{3RR}	not P_{3R1} (66.67)
P_R	not P_1-P_3 (38.89)	P_{R1}	BMI>23 (50.89)	P_{R11}	CETP_2=CC (64.00)
				P_{R1R}	not P_{R11} (47.13)
		P_{RR}	not P_{R1} (27.87)	P_{RR1}	AGT_3=AA (30.38)
				P_{RR2}	AGT_3=CA (26.32)
				P_{RRR}	not P_{RR1}-P_{RR2} (0.00)

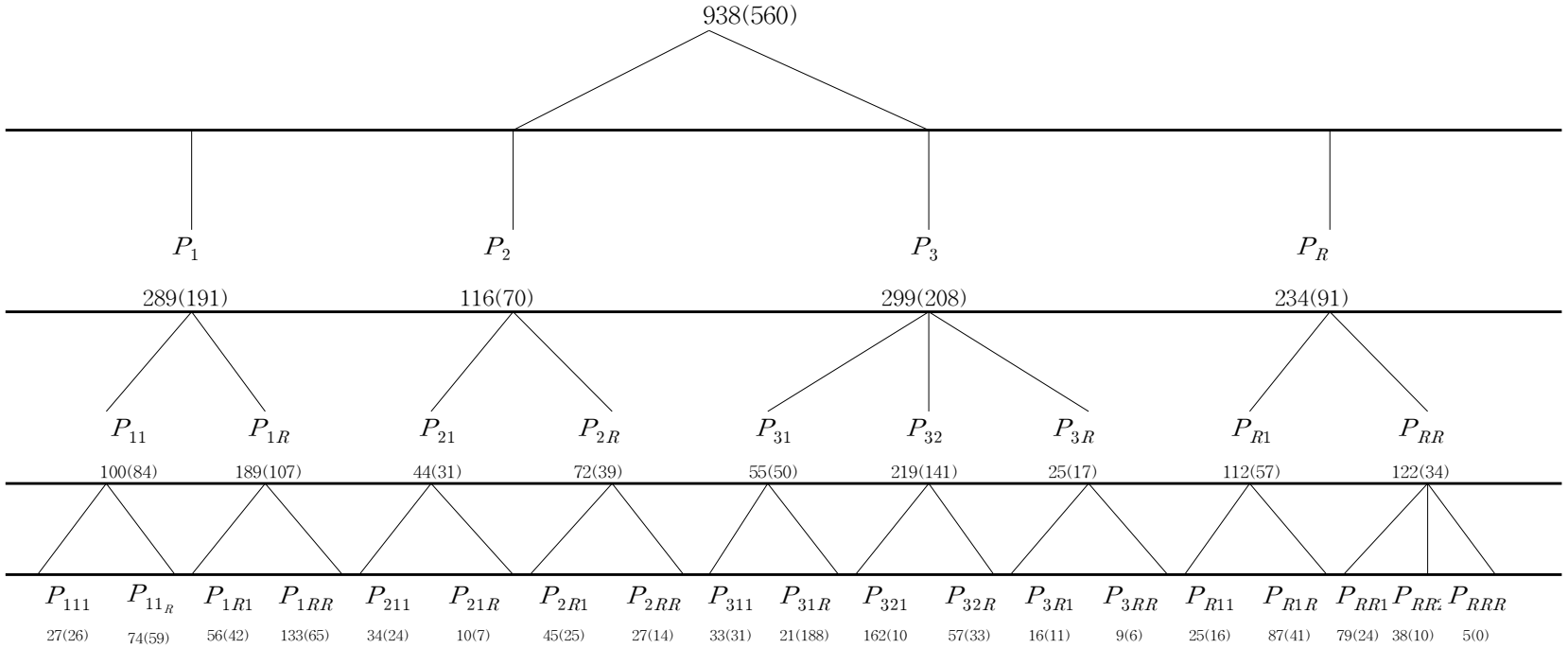

그림 2. 3단계 PRIM에서의 고혈압 위험군 분류

4.5.2 로지스틱 회귀분석

다중 로지스틱 회귀분석에 의한 고혈압 발생 예측 확률은 [표 7의] 결과로 볼 수 있다. 다중 로지스틱 회귀분석은 각각의 경우의 수를 모두 나타내므로 아래의 [표7]는 예측 확률을 순서대로 정리한 표이다.

표 6. 다중 로지스틱 회귀분석에 따른 고혈압 발생 예측 확률

Age>=45	Sex	Drink	Exercise	DM	BMI>23	HDL<40	Tg>=150	AGT_3	CETP_2	LDLR	Prob.(%)
No	M	Yes	No	No	No	No	No	CC	CA	CC	4.4
No	F	Yes	No	No	No	No	No	CC	CA	CC	5.57
No	M	Yes	No	No	No	No	No	AA	CC	CT	10.12
No	M	Yes	Yes	No	No	No	No	CC	CA	CR	11.91
No	F	No	No	No	No	No	No	CC	AA	CC	12.24
No	M	Yes	No	No	No	No	No	AA	CA	CT	12.25
Yes	F	Yes	No	No	No	No	No	CC	CA	CC	12.67
No	F	Yes	No	No	No	No	No	CA	CC	CT	12.86
No	M	Yes	No	No	No	No	No	CA	CC	CC	13.29
...	...	...	...	...	...	...	...	...	...	...	...
Yes	F	No	Yes	No	Yes	Yes	Yes	AA	AA	CC	94.33
Yes	M	Yes	No	Yes	Yes	Yes	No	AA	AA	CC	94.74
Yes	M	Yes	Yes	Yes	No	Yes	Yes	AA	AA	CC	95.41
Yes	M	Yes	Yes	Yes	Yes	Yes	No	CA	CA	CC	95.82
Yes	M	Yes	No	Yes	Yes	Yes	Yes	AA	AA	CC	95.9
Yes	M	Yes	Yes	Yes	Yes	Yes	Yes	AA	CA	CC	96.68
Yes	M	No	Yes	Yes	Yes	Yes	No	AA	CA	CC	96.91
Yes	M	Yes	Yes	Yes	Yes	Yes	No	AA	AA	CC	97.44
Yes	M	Yes	Yes	Yes	Yes	Yes	Yes	AA	AA	CC	98.01

이를 앞서 분석한PRIM의 분할에 적용시켜 보면 아래 [표 8]과 같은 결과를 얻을 수 있다.

로지스틱 회귀 분석을 수행하면 P_1의 경우인 나이 45세 남자의 고혈압 예측 확률은 25.55%에서 98.01%, P_2의 경우인 나이 45세 남자의 고혈압 예측 확률은 43.73%에서 83.53%로 과소추정(under estimation) 또는 과추정(over estimation)되는 경향을 보인다. 또한 로지스틱 회귀분석은 모든 경우의 조합을 보기 때문에 임상적인 해석의 어려움의 한계가 있음을 볼 수 있다.

표 7. PRIM에 따른 다중 로지스틱 회귀분석의 고혈압 발생 예측 확률

PRIM의 partition label			Age ≥ 45	Sex	Drink	Exercise	DM	BMI > 23	HDL < 40	Tg ≥ 150	AGT_3	CETP_2	LDLR	Prob.(%)
P_1	P_{11}	P_{111}	Yes	M	Yes	No	No	Yes	Yes	No	CA	AA	CT	72.27
			Yes	M	Yes	No	No	Yes	Yes	Yes	CA	AA	CT	77.19
			Yes	M	Yes	No	No	Yes	Yes	Yes	AA	AA	CC	81.50
			...	...	...	...	...	...	...	...	...	...	...	...
			Yes	M	Yes	No	Yes	Yes	Yes	Yes	AA	AA	CC	95.90
			Yes	M	Yes	Yes	Yes	Yes	Yes	No	AA	AA	CC	97.44
			Yes	M	Yes	Yes	Yes	Yes	Yes	Yes	AA	AA	CC	98.01
		P_{11R}	Yes	M	Yes	Yes	No	Yes	Yes	No	CC	CA	CT	44.08
			Yes	M	No	Yes	No	Yes	Yes	Yes	CC	CA	CT	58.85
			Yes	M	Yes	No	No	Yes	Yes	No	AA	CA	CT	60.10
			...	...	...	...	...	...	...	...	...	...	...	...
			Yes	M	Yes	Yes	Yes	Yes	Yes	No	CA	CA	CC	95.82
			Yes	M	Yes	Yes	Yes	Yes	Yes	Yes	AA	CA	CC	96.68
			Yes	M	No	Yes	Yes	Yes	Yes	No	AA	CA	CC	96.91
		P_{1R1}	Yes	M	Yes	No	No	No	No	No	CA	AA	CT	37.26
			Yes	M	Yes	No	No	No	No	Yes	CA	AA	CT	43.54
			Yes	M	Yes	No	No	No	No	No	AA	AA	CC	43.60
			...	...	...	...	...	...	...	...	...	...	...	...
			Yes	M	Yes	Yes	No	Yes	No	Yes	CA	AA	CC	83.72

표 7. PRIM에 따른 다중 로지스틱 회귀분석의 고혈압 발생 예측 확률(계속)

PRIM의 partition label			Age ≥ 45	Sex	Drink	Exercise	DM	BMI > 23	HDL < 40	Tg ≥ 150	AGT_3	CETP_2	LDLR	Prob.(%)
			Yes	M	Yes	Yes	Yes	No	Yes	Yes	AA	AA	CT	93.99
			Yes	M	Yes	Yes	Yes	No	Yes	Yes	AA	AA	CC	95.41
		P_{1RR}	Yes	M	Yes	No	No	No	No	No	AA	CA	CT	25.55
			Yes	M	No	No	No	Yes	No	No	CC	CA	CC	27.28
			Yes	M	Yes	No	No	No	No	No	AA	CA	CC	31.36
			...	...	...	...	...	...	...	...	...	...	...	...
			Yes	M	No	Yes	No	Yes	No	No	CA	CA	CC	76.57
			Yes	M	Yes	Yes	Yes	No	No	No	AA	CA	CC	83.65
			Yes	M	Yes	No	Yes	Yes	No	Yes	AA	CA	CC	88.19
P_2	P_{21}	P_{211}	Yes	F	Yes	No	No	No	No	Yes	CA	CA	CC	43.73
			Yes	F	Yes	No	No	No	Yes	Yes	AA	CA	CC	58.43
			Yes	F	Yes	No	No	Yes	No	Yes	AA	CC	CC	59.26
			...	...	...	...	...	...	...	...	...	...	...	...
			Yes	F	Yes	Yes	No	Yes	No	Yes	AA	AA	CC	86.57
			Yes	F	Yes	Yes	No	Yes	Yes	Yes	AA	CA	CC	87.57
			Yes	F	Yes	Yes	No	Yes	Yes	Yes	AA	AA	CC	92.26
		P_{21R}	Yes	F	Yes	Yes	No	No	No	Yes	AA	CC	CT	49.31
			Yes	F	Yes	Yes	No	No	No	Yes	AA	CA	CT	54.67
			Yes	F	Yes	No	No	Yes	No	Yes	AA	CA	CT	57.55

표 7. PRIM에 따른 다중 로지스틱 회귀분석의 고혈압 발생 예측 확률(계속)

PRIM의 partition label			Age ≥ 45	Sex	Drink	Exercise	DM	BMI > 23	HDL < 40	Tg ≥ 150	AGT_3	CETP_2	LDLR	Prob.(%)
			...	...	...	...	...	...	...	...	...	...	...	...
			Yes	F	Yes	No	No	Yes	No	Yes	AA	AA	CT	69.64
			Yes	F	Yes	Yes	No	Yes	No	Yes	AA	CA	CT	74.11
			Yes	F	Yes	Yes	No	Yes	No	Yes	AA	AA	CT	82.89
	P_{2R}	P_{2R1}	Yes	F	Yes	No	No	No	No	No	AA	CC	CC	32.07
			Yes	F	Yes	No	No	No	No	No	AA	CA	CC	36.92
			Yes	F	Yes	No	No	Yes	No	No	AA	CC	CT	45.70
			...	...	...	...	...	...	...	...	...	...	...	...
			Yes	F	Yes	Yes	No	Yes	No	No	AA	CC	CC	70.28
			Yes	F	Yes	Yes	No	Yes	No	No	AA	CA	CC	74.57
			Yes	F	Yes	Yes	No	Yes	Yes	No	AA	CA	CC	84.43
		P_{2RR}	Yes	F	Yes	No	No	No	No	No	CA	CA	CT	31.01
			Yes	F	Yes	No	No	No	No	No	CA	CA	CC	37.43
			Yes	F	Yes	Yes	No	No	No	No	CA	CA	CC	55.81
			...	...	...	...	...	...	...	...	...	...	...	...
			Yes	F	Yes	Yes	No	Yes	Yes	No	CA	CA	CT	80.64
			Yes	F	Yes	No	No	Yes	Yes	No	CA	AA	CC	81.63
			Yes	F	Yes	Yes	No	Yes	No	No	CA	AA	CC	83.53
P_3	P_{31}	P_{311}	Yes	F	No	No	No	Yes	Yes	No	AA	CC	CC	74.32

표 7. PRIM에 따른 다중 로지스틱 회귀분석의 고혈압 발생 예측 확률(계속)

PRIM의 partition label		Age ≥ 45	Sex	Drink	Exercise	DM	BMI > 23	HDL < 40	Tg ≥ 150	AGT_3	CETP_2	LDLR	Prob.(%)
		Yes	F	No	No	No	Yes	Yes	No	AA	CA	CC	78.20
		Yes	F	No	No	No	Yes	Yes	Yes	AA	CC	CC	78.98
		...	...	...	...	...	...	...	...	...	...	...	...
		Yes	F	No	Yes	No	Yes	Yes	Yes	AA	CA	CC	90.77
		Yes	F	No	Yes	No	Yes	Yes	No	AA	AA	CC	92.76
		Yes	F	No	Yes	No	Yes	Yes	Yes	AA	AA	CC	94.33
	P_{31R}	Yes	F	No	No	No	Yes	Yes	No	CA	CA	CT	73.37
		Yes	F	No	No	No	Yes	Yes	Yes	AA	CC	CT	73.85
		Yes	F	No	Yes	No	Yes	Yes	Yes	CC	AA	CT	75.61
		...	...	...	...	...	...	...	...	...	...	...	...
		Yes	F	No	Yes	No	Yes	Yes	Yes	CA	CA	CT	88.31
		Yes	F	No	Yes	No	Yes	Yes	No	CA	CA	CC	88.56
		Yes	F	No	Yes	No	Yes	Yes	Yes	CA	CA	CC	90.96
P_{32}	P_{321}	Yes	F	No	No	No	No	No	No	AA	CC	CC	39.73
		Yes	F	No	No	No	No	No	No	AA	CA	CC	44.98
		Yes	F	No	Yes	No	Yes	No	No	CC	CC	CC	45.01
		...	...	...	...	...	...	...	...	...	...	...	...
		Yes	F	No	Yes	No	Yes	No	No	CA	AA	CC	87.63
		Yes	F	No	Yes	No	Yes	No	Yes	AA	AA	CC	90.00

표 7. PRIM에 따른 다중 로지스틱 회귀분석의 고혈압 발생 예측 확률(계속)

PRIM의 partition label		Age ≥ 45	Sex	Drink	Exercise	DM	BMI > 23	HDL < 40	Tg ≥ 150	AGT_3	CETP_2	LDLR	Prob.(%)
		Yes	F	No	Yes	No	Yes	No	Yes	CA	AA	CC	90.20
	P_{32R}	Yes	F	No	No	No	No	No	No	AA	CA	CT	38.05
		Yes	F	No	No	No	No	No	No	AA	AA	CT	50.97
		Yes	F	No	Yes	No	No	No	No	AA	CC	CT	51.12
		...	...	...	...	...	...	...	...	...	...	...	...
		Yes	F	No	Yes	No	No	Yes	No	AA	AA	CC	84.38
		Yes	F	No	Yes	No	Yes	No	Yes	AA	AA	CT	87.12
		Yes	F	No	Yes	No	Yes	No	Yes	CA	AA	CT	87.37
P_{3R}	P_{3R1}	Yes	F	No	No	No	No	Yes	No	AA	CA	CT	53.18
		Yes	F	No	No	No	No	Yes	No	AA	CC	CC	54.94
		Yes	F	No	No	No	No	Yes	Yes	AA	CA	CT	59.60
		...	...	...	...	...	...	...	...	...	...	...	...
		Yes	F	No	No	No	No	Yes	Yes	AA	AA	CC	76.87
		Yes	F	No	Yes	No	No	Yes	Yes	AA	AA	CT	84.06
		Yes	F	No	Yes	No	No	Yes	No	AA	AA	CC	84.38
	P_{3RR}	Yes	F	No	No	No	No	Yes	Yes	CA	CA	CC	66.74
		Yes	F	No	No	No	No	Yes	No	CA	AA	CC	72.34
		Yes	F	No	Yes	No	No	Yes	Yes	CA	CA	CT	76.10
		...	...	...	...	...	...	...	...	...	...	...	...

표 7. PRIM에 따른 다중 로지스틱 회귀분석의 고혈압 발생 예측 확률(계속)

PRIM의 partition label			Age ≥ 45	Sex	Drink	Exercise	DM	BMI > 23	HDL < 40	Tg ≥ 150	AGT_3	CETP_2	LDLR	Prob.(%)
			Yes	F	No	Yes	No	No	Yes	No	CA	CA	CC	76.54
			Yes	F	No	Yes	No	No	Yes	Yes	CA	CA	CC	80.91
P_R	P_{R1}	P_{R11}	No	M	Yes	No	No	Yes	No	No	CA	CC	CC	26.67
			No	M	Yes	No	No	Yes	No	Yes	CA	CC	CC	32.08
			No	M	Yes	Yes	No	Yes	No	No	AA	CC	CT	36.08
			...	...	...	...	...	...	...	...	...	...	...	...
			No	M	Yes	Yes	No	Yes	Yes	Yes	AA	CC	CC	64.34
			No	M	No	Yes	No	Yes	Yes	No	CA	CC	CC	66.48
			No	F	No	Yes	No	Yes	Yes	Yes	AA	CC	CC	76.35
		P_{R1R}	No	M	Yes	No	No	Yes	Yes	Yes	CC	CA	CC	20.80
			No	F	Yes	No	No	Yes	No	No	AA	CA	CT	29.81
			No	M	Yes	No	No	Yes	No	No	AA	CA	CC	30.61
			No	M	No	No	No	Yes	No	No	AA	CA	CT	31.64
			...	...	...	...	...	...	...	...	...	...	...	...
			No	M	Yes	Yes	No	Yes	Yes	No	AA	AA	CC	74.45
			No	F	No	Yes	No	Yes	No	Yes	AA	AA	CC	78.55
			No	M	Yes	Yes	No	Yes	Yes	Yes	AA	AA	CC	79.10
	P_{RR}	P_{RR1}	No	M	Yes	No	No	No	No	No	AA	CA	CT	12.25
			No	M	Yes	No	No	No	No	No	AA	CA	CC	15.67

표 7. PRIM에 따른 다중 로지스틱 회귀분석의 고혈압 발생 예측 확률(계속)

PRIM의 partition label	Age ≥ 45	Sex	Drink	Exercise	DM	BMI > 23	HDL < 40	Tg ≥ 150	AGT_3	CETP_2	LDLR	Prob.(%)
	No	F	Yes	No	No	No	No	No	AA	CC	CC	16.11
	...	...	...	...	...	...	...	...	...	...	...	...
	No	F	No	Yes	No	No	No	No	AA	AA	CT	47.17
	No	F	No	Yes	No	No	Yes	Yes	AA	CA	CC	62.78
	No	M	No	Yes	No	No	Yes	Yes	AA	AA	CC	69.02
P_{RR2}	No	F	Yes	No	No	No	No	No	CA	CC	CC	16.41
	No	M	Yes	No	No	No	No	No	CA	AA	CC	24.33
	No	M	No	No	No	No	No	No	CA	AA	CT	25.23
	...	...	...	...	...	...	...	...	...	...	...	...
	No	F	No	Yes	No	No	No	No	CA	CA	CC	41.79
	No	F	Yes	No	No	No	Yes	No	CA	AA	CC	43.24
	No	M	No	Yes	No	No	Yes	No	CA	AA	CT	56.85

제 5장 결론 및 고찰

지금까지 심혈관유전체 연구센터의 자료를 통한 고혈압 위험군을 분류하기 위하여 PRIM방법을 사용하였으며, 그 결과를 기존에 분류방법으로 흔히 사용하는 로지스틱 회귀분석의 결과와 비교하였다.

본 논문에서는 고혈압을 진단받은 환자 560명과 정상대조군 378명, 총 938명을 대상으로 고혈압 위험군에는 최종적으로 어떠한 SNP를 가지고 있는지 구체적인 분류를 위하여 연구하였다. 즉, PRIM을 적용하여 고혈압의 위험군을 분류하기 위하여 기존에 고혈압의 위험요인으로 잘 알려진 나이, 성별, 음주유무, 운동유무, 당뇨진단 여부 5가지의 변수로 위험군 3군과 나머지 군 즉, 4개의 집단으로 분류하였고, 고혈압 발생의 위험요소로 알려진 HDL, Tg, BMI 세 변수로 첫 번째에 나누어진 4개의 집단에 PRIM을 각각 적용하여 세부적 위험군을 분류하였다. 여기서 우리는 기존에 알려진 전통적인 위험 요인에 지질변수인 HDL과 Tg, 그리고 BMI 변수를 추가함에 따라 고혈압 위험군에 위험 확률이 더 가중되어 증가됨을 볼 수 있었다. 마지막으로 우리의 관심 요소인 유전적인 요인을 세 번째 단계 PRIM에 적용시킨 결과 두 번째 단계의 위험군의 위험 확률에 유전적 위험확률을 추가한 최종적인 위험군을 구체화시켜 분리 할 수 있었다.

그 결과 PRIM 방법에서 분류된 45세 남자, BMI 23kg/m^2초과, HDL 40mg/dl 미만이면서 동시에 CEPT_2 AA의 SNP를 갖는 분류가 27명 중 26명이 고혈압 환자로 분류되어 96.30%의 최고 위험군으로 분류되었다. 이와 같은 PRIM의 분류기준에 따라 다중 로지스틱 회귀분석을 적용하여 고혈압 위험군을 분류한 결과 72.27%에서 98.01%로 위험 확률이 추정되었다. 두 번째로 위험군으로 분류된 집단은 나이 45세 이상 여자, 비음주와 당뇨변 진단을 받지 않았으며, BMI 23kg/m^2초과, HDL 40mg/dl미만이면서 동시에 AGT_3의 AA와 동시에 LDLR의 CC를 갖는다. 이 집단은 33명중 31명이 고혈압 환자로 분류되어 93.94%의 위험군으로 분류되었다. 마찬가지로 이 분류의 기준에 따라 다중 로지스틱 회귀분석을 적용하

여 고혈압 위험군을 분류한 결과 74.32%에서 94.3%로 추정되었다. 이들 결과는 모두 실제 분류 결과보다 과소 추정(under estimation) 또는 과추정(over estimation)되는 경향을 보인다.

다중 로지스틱 회귀 분석은 각각의 경우의 확률을 모두 볼 수 있으나 임상적 해석상의 어려움을 보인다. 이에 반해 PRIM은 기존의 분류 방법과는 달리 위험 요인들을 가중시켜 고위험 군을 찾아낼 수 있으며 현실적으로 자료를 분석하는 데 있어서 임상적으로 의미 있는 해석이 가능한 장점이 있다. 하지만 추출된 각각의 상자가 겹쳐지는 정도를 기준으로 여러 개의 군집을 하나의 군집으로 묶을 수 있기 때문에 데이터에 대한 사용자의 지식을 기반으로 한 판단이 요구된다.

참 고 문 헌

이혜진, 2단계 모형을 이용한 관상동맥질환 발생 위험에 미치는 유전적 기여도 예측, 2007

하정윤, MDR을 이용한 간경변증 질환 발생의 위험군 분류, 2007

강명욱, 김영일, 안철환, 이용구, 회귀분석, 율곡출판사, 1997

이재원, 박미라, 유한나, 생명과학연구를 위한 통계적 방법, 자유아카데미, 2005

허문열, 이승천, 차경준, 박종선, 유종영, R과 통계계산, 박영사, 2005

Andreas Krause, Melvin Olson, The Basic of S-PLUS, Springer, 2005

David G. Kleinbaum Mitchel Klein, Logistic Regression, Springer, 2002

David W. Hosmer, Stanley Lemeshow. Applied Logistic Regression. Wiley, 2002

Hastie,. Tibshirani., Friedman. The Elements of Statistical Learning. Springer, 2001

W.N. Venable, B.D. Ripley. Modern Applied Statistics with S-PLUS, Springer, 1999

Friedman, J. H.. and Fisher, N. I. Bump Hunting in High-Dimensional Data.

Computing and Statistics 9, 1999:123-143

Greg Dyson, Ruth Frikke-Schmidt, Børge G. Nordestgaard, Anne Tybjærg-Hansen, Charles F. Sing, An Application of the Patient Rule-Induction Method for Evaluating the Contribution of the *Apolipoprotein E* and *Lipoprotein Lipase Genes* to Predicting Ischemic Heart Disease, *Genet Epidemiol*,. 2007:9;31(6):515-27

Hyndman, R. J. Computing and graphing highest density regions. The *American Statistician*,, 1996:50,120 - 126.

Nannings B, Abu-Hanna A, de Jonge E. Applying PRIM (Patient Rule Induction Method) and logistic regression for selecting high-risk subgroups in very elderly ICU patients. *Int J Med Inform*. 2007:7;21,2408-8

Peter C. Austin, A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality, *Stat Med*. 2007:7;10;26(15):2937-57

ABSTRACT

A classification of genetic high risk group for essential hypertension using PRIM

Ahn, Jin Hee
Dept. of Biostatistics and Computing
The Graduate School
Yonsei University

The complex disease such as hypertension has become known to be contributed by the different combinations of genetic and environmental risk factors. Therefore, there is a need to use an appropriate method that can identify subgroups of individuals at substantially increased risk of hypertension to be each characterized by a particular combination of risk factors.

In this thesis, we proposed an application of the Patient Rule Induction Method(PRIM) to classify genetic high risk group for hypertension. The PRIM had an advantage that it could identify subgroups and individuals at risk by integrating data from many genetic factors and multiple environmental variables in detail. The PRIM was applied to model the cumulative incidence of hypertension in a sample of 938 unrelated individuals(560 patients with hypertension and 378 normal individuals) with 17 single nucleotide

polymorphisms(SNP) 10 environmental variables from Yonsei Cardiovascular Genome Center.

In the first stage, the traditional risk factors for hypertension such as age, sex, drinking history, regular exercise, and diabetes were analyzed.
Secondly, the PRIM was executed with variables of serum lipid profiles, and body mass index, and in the final stage, it was conducted with multiple SNPs. The result of analysis was compared with that of logistic regression analysis which was commonly used to evaluate and classify risk factors.

In conclusion, we found that the PRIM had some utilities that it could identify the degree of risk in each partition with combination of risk factors and find out the highest risky group. Also, it could find partitions where the risk of hypertension was high and maximize the number of hypertension cases explained by the partitions. So, it could support to draw the implications that were clinically meaningful.

Key words : essential hypertension, patient rule induction method, logistic regression