혼합모형을 이용한 관상동맥질환의 유전적 관련성 분석

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
손 낙 훈

혼합모형을 이용한 관상동맥질환의 유전적 관련성 분석

지도 장 양 수 교수

이 논문을 석사 학위논문으로 제출함

2009년 6월 일

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
손 낙 훈

손낙훈의 석사 학위논문을 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

연세대학교 대학원

2009년 6월 일

감사의 글

한여름의 높은 기온만큼 뜨거운 열정과 장맛비의 굵은 빗방울들만큼이나 많은 땀방울들이 어우러진 석사생활을 마무리 하고자 합니다. 지난 2년 동안 연을 맺었던 모든 분들께 감사드립니다.

먼저 바쁘신 와중에 학위논문 심사를 맡아주신 장양수 교수님, 입학부터 이 논문이 나오기까지 여러 방면으로 지도해주신 송기준 선생님, 먼 곳에서도 여러모로 관심 가져주신 명성민 선생님, 애정과 관심을 가지고 지도해주신 조진남 선생님, 유전통계의 기반을 다져주신 임길섭 선생님, 여러 가지 생각의 기회를 주신 남정모 선생님, 새로운 관심사를 만들어주신 임현선 선생님께 깊은 감사의 마음을 전합니다.

같은 수업을 들으며 많은 것을 배웠던 고희중 선배님, 짧은 학과생활동안 돌봐주신 무영이형, 이것저것 많은 것을 알려주신 원열이형, 종종 들리셔서 상담을 해주셨던 봉섭이형, 늘 웃으며 이야기를 들어주시던 혜리누나, 친 누나처럼 챙겨주셔서 언제나 따르고 싶은 은희누나, 시시콜콜 대화를 나눴던 수희선배, 구박쟁이 영애선배, 동네주민 경화선배, 직속상관 진희선배, 조용히 웃어주던 성유선배, 티격태격 했지만 여러모로 큰 도움을 준 혜선선배, 즐거운 대학원생활을 알려준 용진이까지 든든한 선배님들이 있어서 석사생활을 마무리 할 수 있었습니다.

잘 챙겨주던 마음씨 착한 수연이, 조용한 성희, 독특한 성훈이, 유쾌함과 여유로움을 배우고 싶은 성혁이, 종종 큰 웃음주던 하나, 애교덩이 지연이, 유전체센터에 잘 적응해준 광복이가 있어서 힘이 되었습니다.

대학원 생활에 소홀했던 평생지기 친구들 성재, 도호, 석호, 진현, 준호, 우석에게 미안한 마음을 전하고, 학부의 인연으로 서로 걱정해주는 건희, 혜지, 선영이, 혜란이, 소중한 학과인연들 효경이, 철유, 경태, 동주, 강현이, 동훈이, 대곤이, 일석이, 승민이형, 학원에서 만나 서로 잘되길 바라는 정민이, 말하지 않아도 통하는 abc, 인희. 소소한 일상으로 즐거움을 주는 지연이가 있어서 외롭지 않았습니다.

유전체센터의 신동직 선생님, 든든한 카운셀러 진우 선생님, 최고로 잘 챙겨주시

는 은정 선생님, 끊임없는 일 요청에도 모두 응해주신 계원, 윤희, 상희 선생님, 종알종알 현정씨, 골목대장 뒷태 아람이, 부끄럼쟁이 소연이, 고마운 지혜 선생님이 있어서 유전체센터 생활이 행복했습니다.

끝으로 세상에 그 무엇과도 바꿀 수 없는 사랑하는 우리가족 할아버지, 할머니, 아버지, 어머니, 유경이에게 사랑의 마음을 전합니다. 앞으로 더 큰 사람이 되기 위해 노력하겠습니다.

2009년 6월

손낙훈 올림

차 례

표 차 례

그 림 차 례

국문요약

혼합모형을 이용한 관상동맥질환의 유전적 관련성 분석

질병발생에 대한 유전적 관련성을 분석할 경우 유전자간의 복잡한 교호작용과 유전자와 환경적 요인과의 교호작용을 고려하는데, 흔히 사용하는 통계분석방법인 로지스틱 회귀분석에서는 SNP의 숫자가 많아지면 그것들 사이의 복잡한 교호작용 효과에 대한 해석이 어려워지게 된다.

본 논문에서는 유전형조합을 이용하여 많은 수의 SNP간의 교호작용 및 환경적 인자와의 교호작용을 효율적으로 분석할 수 있는 방법으로 혼합모형을 사용하는 것을 제안하였다. 실제자료로 연세대학교 심혈관계질환 유전체연구센터에서 조사된 관상동맥질환 환자군 503명, 정상대조군 503명, 총 1,006명의 대상을 이용하였다. 관심 있는 32종의 SNP중 3개, 4개, 5개의 SNP에 대한 유전형조합을 만들어 임의효과로 정의하였고, 전통적으로 관상동맥질환에 영향을 미치는 임상변수 12종을 고정효과로 정의하였다. 임의효과모형과 혼합효과모형에서 Odds ratio를 비교하였고, 기존의 분석방법인 로지스틱 회귀분석을 적용한 결과와 비교하였다. 혼합모형을 적용한 결과 임의효과를 통해 모집단을 대표한 개별특성을 고려한 분석을 시행할 수 있었는데, 관상동맥질환과의 관련성이 유전형조합별로 각기 다르게 나타나고 있음을 알 수 있었다. 결론적으로, 질병발생에 대한 SNP들의 유전형 조합에 따른 개별적 효과를 구체적으로 파악하고자 할 경우, 혼합모형이 유용하게 적용될 수 있음을 확인할 수 있었다.

핵심 되는 말 : 관상동맥질환, 혼합모형, 유전적 관련성

제 1 장 서론

1.1 연구 배경 및 목적

최근 급속한 경제 성장과 생활양식의 서구화는 국내의 질병 양상에 많은 변화를 가져왔다. 고지방 식품의 섭취량 증가, 스트레스 증가, 운동부족, 의료 기술 발달에 의한 노령 인구의 증가 등으로 인하여 심혈관계질환의 발생률, 유병률 및 사망률이 급격하게 증가하고 있다(한국인 질환유전자 발굴 연구에 관한 보고서 2008). 특히 한국인은 고지방 식사에 취약한 유전배경을 가지고 있어, 심혈관계질환의 발생과 관련된 생활습관이나 환경적요인과 더불어 유전적 위험요인의 중요성이 부각되고 있다. 따라서 유전배경을 밝히기 위하여 단일 염기 다형성(single nucleotide polymorphism; SNP, 이하 SNP)자료 분석에 많은 연구가 진행 중이다.

질병발생에 대한 유전적 관련성을 분석할 경우, 유전자간의 복잡한 교호작용(interaction)과 유전자와 환경적 요인과의 교호작용을 고려하여 진행하는데 흔히 사용하는 통계분석 방법인 로지스틱 회귀분석에서는 고려하고자 하는 SNP의 숫자가 많아지면 그것들 사이의 복잡한 교호작용의 영향을 검증하고 해석하는데 난해한 단점이 있다(Foulkes, 2005). 따라서 많은 수의 SNP간의 교호작용 및 환경적 인자와의 교호작용을 효율적으로 분석할 수 있는 방법으로 혼합모형을 사용하여 분석하고자 한다.

지금까지 SNP자료의 분석에 여러 가지 방법이 제안되었는데 장단점은 다음과 같다. 로지스틱회귀분석방법은 잠재적인 환경적인자의 영향을 통제하고 개별 SNP의 효과를 밝힐 수 있는 장점을 가지고 있다. 하지만 고차원의 SNP간의 교호작용 및 환경적 인자와의 교호작용을 분석하기에는 어려움이 따른다. 관심 SNP수가 증가함에 따라 유전형의 개수는 3^{SNPs}수만큼씩 증가하여 주어진 표본 크기 내에서 분석이 불가능해질 수도 있다는 단점이 있다(Foulkes, 2005). 반복분할(recursive partitioning)방법은 회귀분석방법에서 얻을 수 없는 고차원의 교호작용을 확인 할 수 있고, 비모수방법으로의 적용도 가능한 장점이 있다. 하지만 자료구조에 제약을 받는 단점이 있다(Breiman, 1984). 조합분할(combinatorial partitioning)방법은 반복분할방법에서 제곱합의 개념을

추가한 것으로 유전자형(genotype)의 다중조합(multiple combination)을 고려할 수 있다는 장점이 있다. 단점으로는 통계적 유의성 평가 방법으로 순열검정이 제안되어, 3개 이상의 SNP의 상호작용을 고려한 분석에는 계산이 복잡해진다(Nelson, 2001). 순열검정(permutation test)방법은 질병과 유전자의 관계연구나 Microarray와 같은 고차원의 자료에 적용 할 수 있다. 많은 수의 후보 SNP를 효과적으로 줄일 수 있는 장점이 있지만, 공변량의 조절이나 유전자와 환경요인의 교호작용에 대한 제한점이 있다(Hoh, 2000).

1.2 연구 내용 및 방법

앞에 나열한 여러 방법들의 단점을 보완하고자 유전형집단(genotype group)을 생성하여 여러 개의 개별 SNP의 효과가 아닌 SNP조합을 동시에 고려해 유전형집단과 관상동맥질환의 관련성을 보고자 한다. 이때, 유전형집단을 임의효과로 설정하여 유전형조합에 의한 변동량을 고려한 혼합모형에 적용시켜 보고자 한다.

1.3 논문의 구성

1장에서는 연구의 배경 및 목적에 대해 소개하고 연구 내용 및 방법에 대해 제시한다. 2장에서는 연구 방법으로 일반화 선형 혼합 모형의 이론과 개념, 추정방법, 검정방법, 적합도 평가방법에 대해 소개한다. 3장에서는 실제자료를 이용하여 관상동맥질환의 유전적 관련성 분석을 실시한다. 4장에서는 결론 및 고찰에 관하여 논의한다.

제 2 장 이론적 배경

2.1 일반화 선형 혼합 모형의 정의 및 가정

일반적으로 범주형 자료를 분석하기 위해서는 분산을 안정시키기 위하여 변환된 자료를 이용하여 선형 모형을 적합 시키는 방법 및 일반화 선형 모형 등이 적용되어져왔다(이준영, 1999). 일반화 선형 모형은 반응값들의 평균에 대한 비선형 함수를 모형화 할 수 있고, 많은 경우에 자료의 분포에 대해 보다 직접적인 모형화가 가능하다. 특히 자료의 분포에 대한 직접적인 모형화를 할 수 있다는 점은 범주형 자료의 경우, 실험자가 표본 계획을 통해 자료의 분포를 조절 할 수 있다는 면에서 특별한 중요성을 지닌다. 일반화 선형 혼합 모형은 일반화 선형 모형과 혼합 모형을 연결함으로서 범주형 자료중 집락자료, 상관자료, 반복 측정된 자료에 대한 과산포 문제를 다룰 수 있게 해주었다.

일반화 선형 혼합 모형의 식은

$$y = X\beta + Z\nu + \epsilon$$

으로 나타낼 수 있다. 여기서 X는 고정효과를 나타내는 계획행렬, β는 고정효과 벡터, Z는 임의효과를 나타내는 계획행렬, ν는 평균이 0이고 분산이 G인 다변량 정규분포를 따르는 임의효과 벡터, ϵ는 평균이 0이고 분산이 R인 다변량 정규분포를 따르는 오차 벡터이다. 일반화 선형 혼합 모형을 사용하기 위해서는, 첫째, 임의효과가 주어진 상태에서 반응값들의 분포(주로 지수족 분포를 가정한다)를 가정하고, 둘째, 고정효과와 임의효과를 포함하고 있는 선형추정량을 정의해야 하며, 셋째, 반응값들의 조건부 평균과 선형추정량을 연결해 주는 연결함수를 설정해야하고, 넷째, 임의효과에 대한 분포(정규분포를 가정한다)를 가정해야 한다.

2.2 일반화 선형 혼합 모형의 모수추정

임의효과가 주어진 상태에서 반응값 y에 대한 확률밀도함수는

$$f_{y|\nu}(y|\nu,\beta,w)=\exp\left[\frac{1}{a(w)}(y\theta-c(\theta))+d(y,w)\right]$$

이다. 이때, θ는 정준모수이고, w는 알려진 가중값을 가정하고, $\mu=E[y|\nu]$, 연결함수를 $g(\bullet)$으로 표현한다면, 일반화된 선형 혼합 모형에 의해 조건부 평균(μ)와 선형추정량(η)은

$$g(\mu)=\eta=X\beta+Z\nu$$

과 같은 관계를 가진다.

여기서 임의효과에 대해 정규성 가정을 사용하는 주된 이유는 임의효과들 사이의 복잡한 공분산구조에 대한 설정이 간편해지기 때문이다(이준영, 2000). 또한, 가중값 w는 해석상의 편의를 위해 가정한다.

2.2.1 모수 추정 방법

추정 방법으로는 최대우도를 이용한 방법으로

$$\left\{{}_c tr(V^{-1}Z_iZ_i')\right\}_{i=0}^{r}=\left\{{}_c y'PZ_iZ_i'Py\right\}_{i=1}^{r}$$

$$\log l_p(V)=-\frac{1}{2}y'Py-\frac{1}{2}\log|V|-\frac{N}{2}\log(2\pi)$$

$$(P=K(K'VK)^{-1}K')$$

를 통해 모수를 추정하는데, 여기에서 K는 고정효과의 벡터와의 곱이 0이 되는 행렬, V는 모형 전체의 분산을 말한다. 이는 수리적 계산이 어렵고 편향이 존재할 수 있어서 통상적으로는 제한된 최대 우도함수(restricted maximum likelihood; REML)를 사용하여 불편 추정치를 얻게 된다. 제한된 최대 우도함수는

$$K'y \sim N(0, K'VK)$$
$$\{{}_c tr(PZ_iZ_i')\}_{i=0}^{r} = \{{}_c y'PZ_iZ_i'Py\}_{i=0}^{r}$$
$$(P = K(K'VK)^{-1}K')$$

을 통해 추정할 수 있다. 그 외에 추정방법으로는 MINQUE(minimum norm quadratic unbiased estimation)등이 있는데 이는 분산을 미리 정해놓은 상태로 추정하는 것을 기초로 한다.

2.2.1.1 우도적 접근

모수 추정을 위한 우도적 접근으로 우도함수는

$$L = \int \prod_i^n f_{Y_i|\nu}(y_i|\nu) f_\nu(\nu) d\nu$$
$$l = \log \int f_{Y|\nu}(y|\nu) f_\nu(\nu) d\nu = \log f_Y(y)$$

를 정의한다. 고정효과의 추정을 위해 우도함수를 미분하는 방법으로

$$\frac{\partial l}{\partial \beta} = \int \frac{\partial \log f_{Y|\nu}(y|\nu) f_\nu(\nu) d\nu}{\partial \beta} f_{Y|\nu}(y|\nu) f_\nu(\nu) d\nu / f_Y(y)$$
$$= \int \frac{\partial \log f_{Y|\nu}(y|\nu)}{\partial \beta} f_{\nu|y}(\nu|y) d\nu$$

$$\frac{\partial l}{\partial \beta} = \int X' W^*(y-\mu) f_{\nu|y}(\nu|y) d\nu$$

$$= X' E[W^*|y] - X' E[W^* \mu|y]$$

$$(W^* = diag\{[a(\theta) VAR(\mu_i) g_\mu(\mu_i)]^{-1}\})$$

고정효과를 추정할 수 있다. 같은 방법으로 임의효과도 우도함수의 미분을 통해

$$\frac{\partial l}{\partial \nu^*} = \int \frac{\partial \log f_\nu(\nu)}{\partial \nu^*} f_{\nu|y}(\nu|y) d\nu$$

$$= E\left[\frac{\partial \log f_\nu(\nu)}{\partial \nu^*}|y\right]$$

추정할 수 있다. 하지만 이 방법을 직접적으로 적용시키는 것은 수리적 반복 계산이 복잡하여 실제 자료에 적용시키기는 어렵다.

2.2.1.2 주변 우도적 접근

주변 우도(marginal likelihood)적 접근에서 보면, 관찰값들의 결합 우도 대신 임의효과에 대한 적분을 통해서 임의효과를 제거한 주변 우도함수를 이용하여 모수 β와 G에 해당하는 모수 Σ의 최대 우도 추정값을 얻을 수 있다. 이때 모수 β와 모수 Σ의 최대 우도 추정값인 $\hat{\beta}$과 $\hat{\Sigma}$은 다음의 주변 우도 함수

$$l(\beta, \Sigma|y) = \int \prod_{i=1}^{n} f(y_i|\nu, \beta, w_i) f(\nu|\Sigma) d\nu$$

를 최대화함으로서 얻어진다. 또한, 그룹간 임의효과들이 서로 독립인 경우에 주변 우도 함수는

$$l(\beta,\Sigma|y)=\prod_{i=1}^{t}\int\prod_{j=1}^{n_i}f(y_{ij}|\nu_i,\beta,w_{ij})f(\nu_i|\Sigma)d\nu_i$$

로 표현할 수 있다. 이때 적분의 차원은 주어진 그룹 내에서 몇 개의 임의효과가 있는가에 따라 결정된다. 적분값내 결합 밀도 함수의 곱의 개수는 집락내 관찰값의 수에 따라 비례적 증가한다. 하지만, 실제 적분이 어려워 의사 우도 함수를 이용하거나, 모의실험, 뉴튼-랩슨방법을 이용해 수치적 근사방법을 사용하여 근사값을 얻고, 수치적으로 최대화 시켜 모수의 추정값을 얻게 된다(Breslow, 1993).

2.2.1.3 조건부 우도적 접근

조건부 우도적 접근 방법은 자료에서 주어진 임의효과의 조건부분포를 이용하는 방법으로서 예를 들어 ν_i가 주어진 상황에서 y_{ij}가 독립이며 $\pi(x_{ij})$인 베르누이분포를 따르는 경우

$$P(y_{i1}=1,y_{i2}=0|S_i=1)=\frac{1}{1+e^{-\beta}} \quad ,\text{where } S_i=y_i\,.$$

의 조건부 확률을 통해 추정값을 얻을 수 있다. 하지만 분포가정의 문제에서 정보력을 상실하는 단점과, 임의효과들의 비교를 통해 얻는 정보력의 상실 등으로 인해 임의효과에 초점을 두고 있는 경우 사용이 제한적이다.

2.2.1.4 벌점 의사 우도적 접근

의사 우도적 접근방법은 일반적으로 분포가정이 없어 모수추정방법에 있어서 효율적이라는 장점이 있다. Breslow와 Clayton의 벌점 의사 우도(penalized quasi-likelihood;

PQL)을 이용한 적합 방법은 연결 함수로 정준 연결(canonical link)을 가정하는데, 이 때 $\eta_i = \theta_i$가 된다. 반응값들에 대한 결합 밀도 함수를 통해 주변 우도 함수는

$$l(\beta, \Sigma|y) \approx |\Sigma|^{-\frac{1}{2}} \int \exp^{-k(\nu)} d\nu$$

$$k(\nu) = \sum_{i=1}^{n} \frac{(y_i \eta_i - c(\eta_i))}{a(w_i)} + \frac{1}{2} \nu' \Sigma^{-1} \nu$$

로 표현된다. 앞에서 $\frac{d\eta_i}{d\nu} = z_i$ 이므로

$$k'(\nu) = -\sum_{i=1}^{n} \frac{(y_i - c'(\eta_i)) z_i}{a(w_i)} + \Sigma^{-1} \nu$$

$$k''(\nu) = \sum_{i=1}^{n} \frac{c''(\eta_i) z_i z'_i}{a(w_i)} + \Sigma^{-1} = Z' W Z + \Sigma^{-1}$$

가 된다. Z'는 임의효과의 계획행렬이고, W는 i번째 대각원소가 $\frac{c''(\eta_i)}{a(w_i)}$인 대각행렬이다. W의 원소들이 일반화 선형 혼합 모형의 적합을 위해 반복 가중된 제곱 추정값(iteratively weighted least squares estimate)을 얻기 위해 사용되는 가중값이다. Talyor 확장에 의해 $k(\nu)$는

$$k(\nu) = k(\tilde{\nu}) + (\nu - \tilde{\nu})' k'(\tilde{\nu}) + \frac{1}{2} (\nu - \tilde{\nu})' k''(\tilde{\nu}) (\nu - \tilde{\nu}) + O(\| \nu - \tilde{\nu} \|)$$

로 표현되며, $k'(\tilde{\nu}) = 0$인 $\tilde{\nu}$에서 주변 우도 함수는

$$l(\beta, \Sigma|y) \propto |\Sigma|^{-\frac{1}{2}} \exp^{-k(\tilde{\nu})} \int \exp^{-\frac{1}{2}(\nu - \tilde{\nu})' k''(\tilde{\nu})(\nu - \tilde{\nu})} d\nu$$

의 형태를 가진다. 자료의 실제 주변 로그 우도 함수는

$$L(\beta,\Sigma|y) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\log|k''(\tilde{\nu})| - k(\tilde{\nu})$$
$$= -\frac{1}{2}\log|I + Z'WZ\Sigma| - \sum_{i=1}^{n}\frac{(y_i\eta_i - c(\eta_i))}{a(w_i)} - \frac{1}{2}\tilde{\nu}'\Sigma^{-1}\tilde{\nu}$$

가 된다. 이때 가중치 행렬 W의 원소들이 작은 값을 취한다면 주변 로그 우도 함수에서 첫항이 무시되며 벌점 의사 우도

$$-\sum_{i=1}^{n}\frac{(y_i\eta_i - c(\eta_i))}{a(w_i)} - \frac{1}{2}\nu'\Sigma^{-1}\nu$$

가 얻어진다. 모수 β를 추정하기 위해 β와 ν에 대해 편미분하면,

$$\sum_{i=1}^{n}\frac{(y_i - c'(\eta_i))x'_i}{a(w_i)} = 0$$
$$\sum_{i=1}^{n}\frac{(y_i - c'(\eta_i))z'_i}{a(w_i)} = \Sigma^{-1}\nu$$

이 된다. 여기서 ν의 분산-공분산 행렬$\Sigma = \Sigma(\sigma^2)$의 원소들인 분산 성분 벡터 σ^2에 의존하므로, 이를 추정하기 위해 Patterson과 Thompson(1971)의 제한된 최대 우도함수

$$-\frac{1}{2}\log|V| - \frac{1}{2}\log|X'V^{-1}X| - \frac{1}{2}(y - X\hat{\beta})'V^{-1}(y - X\hat{\beta})$$
$$(V = W^{-1} + Z\Sigma Z')$$

를 σ^2으로 편미분 하여 Harville(1977)의 추정방정식에 따라

$$-\frac{1}{2}\left[(y-X\beta)'V^{-1}\frac{\partial V}{\partial\sigma_j^2}V^{-1}(y-X\beta)-tr(P\frac{\partial V}{\partial\sigma_j^2})\right]=0$$

$$(P=V^{-1}-V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1})$$

을 사용한다. 이때 Fisher 정보 행렬은

$$F=\{F_{jk}\}=-\frac{1}{2}tr\left(P\frac{\partial V}{\partial\sigma_j^2}P\frac{\partial V}{\partial\sigma_k^2}\right)$$

이 된다. 즉, Breslow와 Clayton이 제안한 일반화 선형 모형의 적합을 위한 PQL 접근 방식은, 먼저 n개의 자료가 독립이라는 가정하에 일반화 선형 모형 방법을 이용하여 β의 초기추정값 $\hat{\beta}$을 구하고, $\hat{\beta}$을 이용해 잔차를 얻은 뒤, σ^2의 초기추정값 $\widehat{\sigma^2}$을 구해 PQL을 β와 ν에 편미분해 새로운 추정값 $\hat{\beta}$과 $\widehat{\sigma^2}$을 구한후, 이들을 이용하여 Harville의 추정방정식과 Fisher 정보 행렬을 이용해 새로운 추정값 $\widehat{\sigma^2}$을 구해 다시 처음부터 반복해 추정값들이 수렴할 때까지 실시해 원하는 모수들 β와 $\sum$에 대한 최대 우도 추정값을 얻는다.

추정의 다른 방법으로는 최대 로그 유사 우도 추정량(maximum log pseudo-likelihood)

$$l_M(\theta,p)=-\frac{1}{2}\log[V(\theta)]-\frac{1}{2}r'V(\theta)^{-1}r-\frac{f}{2}\log\{2\pi\}$$

$$(r=p-X(X'V^{-1}X)^{-}X'V^{-1}p)$$

을 이용하는 방법과, 제한 로그 유사 우도 추정량(restricted log pseudo-likelihood)

$$l_R(\theta,p)=-\frac{1}{2}\log[V(\theta)]-\frac{1}{2}r'V(\theta)^{-1}r-\frac{1}{2}\log[X'V(\theta)^{-1}X]-\frac{f-k}{2}\log\{2\pi\}$$

$$(r=p-X(X'V^{-1}X)^{-}X'V^{-1}p)$$

을 이용할 수 있다. 여기서 k는 X의 계수이고, θ는 모수이다.

2.2.2 모수 추정을 위한 계산방법

계산방법으로 몇 가지 방법을 소개하면, 첫째, EM 알고리즘은 McLachlan과 Krishnan(1996)의 방법을 기초로 초기값 $\beta^{(0)}$, $\nu^{(0)}$, $D^{(0)}$을 정한 후 최대화 $E\left[\log f_{Y|u}(y|u,\beta,\nu)|y\right]$를 통해 $\beta^{(m+1)}$과 $\nu^{(m+1)}$를, 최대화 $E\left[\log f_u(u|D)|y\right]$를 이용해 $D^{(m+1)}$을 계산하는데, 수렴할 때까지 반복 시행한다. 둘째, 뉴튼-랩슨(Newton-Raphson)방법은 비선형식의 최대화방법으로

$$\theta^{(m+1)} = \theta^{(m)} - \left[\frac{\partial^2 f(\theta)}{\partial\theta\partial\theta'}\right]^{-1}|_{\theta=\theta^{(m)}} f'(\theta^{(m)})$$

$$\sigma^{2(m+1)} = \sigma^{2(m)} - \alpha(l_{\sigma^2\sigma^2})^{-1} l_{\sigma^2}, \quad \text{where } 0 < \alpha \leq 1$$

을 통해 계산한다. 특징으로는 국소최대(local maximum)의 이용으로 수렴하지 않는 경우가 발생하고, 모수공간(parameter space)안에서 반복이 유지될 필요가 없다. 다음으로 수치 구적방법(numerical quadrature)은 일차원 적분형

$$\int_{-\infty}^{\infty} h(u)\frac{e^{-u^2/(2\sigma_u^2)}}{\sqrt{2\pi\sigma_u^2}}du$$

을 이용하는데, 범위가 한정되지 않은 수리적 적분이 어렵기 때문에 평활함수(smooth function)를 이용한 적분방법인 Gauss-Hermite 구적방법

$$\int_{-\infty}^{\infty} e^{2xa-a^2}e^{-x^2}dx$$

을 통해 계산한다. 이 방법은 주변 우도함수에서 하나의 임의효과나 여러 개의 독립적인 임의효과가 존재하는 경우 사용한다.

이 외에도 마르코프 연쇄 몬테칼로 알고리즘(Markov chain Monte Carlo)이나, 확률근사(stochastic approximation) 알고리즘 등의 방법으로도 계산이 가능하다.

2.3 일반화 선형 혼합 모형의 모수 검정

앞서 제안한 방법으로 계산한 추정값들은 F나 t분포를 기초로 검정할 수 있다. 검정통계량을 대비를 이용해 정의해보면, 고정효과의 $C = l(\hat{\beta}) = 0$와 임의효과의 $C = l(\hat{\nu}) = 0$가 되고, F검정을 통해 귀무가설인 $C = 0$을 왈드 통계량(Ward statistic)으로 계산하며,

$$
\begin{aligned}
W_\beta &= (l\hat{\beta})'(VAR(l\hat{\beta}))^{-1}(l\hat{\beta}) \\
&= (l\hat{\beta})'(l\,VAR(\hat{\beta})\,l)^{-1}(l\hat{\beta}) \\
W_\nu &= (l\hat{\nu})'(VAR(l\hat{\nu}))^{-1}(l\hat{\nu}) \\
&= (l\hat{\nu})'(l\,VAR(\hat{\nu})l)^{-1}(l\hat{\nu})
\end{aligned}
$$

을 이용한다. 근사적으로, W는 자유도 l의 카이제곱분포를 따르게 되고, 왈드 F 검정통계량은

$$
F_{df1,\,df2} = \frac{W}{df1}
$$

을 통해 계산할 수 있다. 여기서 $df1$은 대비의 자유도이고, $df2$는 대비 분산의 자유도이다. 만약 대비가 하나만 존재한다면, $t_{df2} = (F_{1,df2})^{1/2} = W^{1/2}$ 이므로 t검정이 가능하다. 이 때, 통계학적 의사결정은 $W > \chi^2_{1,1-\alpha}$이면 유의수준 α에서 귀무가설을 기

각하게 되고, $W \leq \chi^2_{1,1-\alpha}$이면 유의수준 α에서 귀무가설을 채택하게 된다.

2.4 일반화 선형 혼합 모형의 적합도 평가

일반화 선형 모형의 적합도를 평가하기 위한 방법의 피어슨 카이제곱통계량은

$$\chi^2 = \sum_i \frac{w_i(y_i - \hat{\mu}_i)^2}{a(\hat{\mu}_i)}$$

이다. 여기서 $\alpha(\hat{\mu}_i)$는 추정된 평균의 분산함수이다. 일반화 선형 혼합모형에서는 일반화 피어슨 카이제곱통계량

$$\chi^2_g = \hat{r}' V(\hat{\theta}^*)^{-1} \hat{r}$$

을 사용한다.

다른 방법으로는 -2 log likelihood 검정통계량을 사용한다. 이 검정통계량은 작을수록 좋은 값으로 단순비교가 쉬운 것이 장점이나, 모수의 개수가 증가할수록 검정통계량의 값이 감소하는 경향이 있다. 따라서 이러한 현상을 보정해 주는 방법으로 Akaike's 정보기준(AIC)과 Schwarz's 기준(BIC)

$$AIC = \log(l) - q$$
$$SIC = \log(l) - \frac{q(\log(N-p)}{2}$$

을 사용할 수 있다. 이때, q는 공분산 모수의 개수이고, p는 고정효과의 개수, N은 전체 관측치의 개수이다.

제 3 장 관상동맥질환의 유전적 관련성 분석

3.1 자료에 대한 개요

분석에 사용한 자료는 연세대학교 심혈관계질환 유전체 연구센터에서 실시하는 검진에 동의한 대상자들의 임상자료를 바탕으로 남자 55세 이하, 여자 60세 이하의 Angio type 2 이상의 관상동맥질환이 있는 환자 503명과 그에 따른 성별 및 연령을 고려하여 추출한 대조군 503명, 총 1,006명의 자료를 이용하였다.

심혈관계질환 유전체 연구센터의 자료에는 나이, 성별, 당뇨병 진단여부, 고혈압 진단여부, 흡연력, 체질량지수(body mass index, BMI), 총 콜레스테롤(total cholesterol, Tchol), 고밀도지단백콜레스테롤(high density lipoprotein, HDL), 저밀도지단백콜레스테롤(low density lipoprotein, LDL), 중성지방(triglyceride, Tg), 혈액요소질소(blood urea nitrogen, BUN), 크레아티닌(creatinine), 인슐린(insulin), hsCRP(high-sensitivity c-reactive protein, hsCRP)의 15가지의 전통적으로 관상동맥질환과 관련이 있는 변수들과 adiponectin, lipoprotein-associated phospholipase A2(Lp-PLA2), interleukin 6(IL6), RANTES의 4가지 biomarker, 32개의 SNP들로 이루어져 있다. 분석대상변수를 선택하기 위한 분석방법에는 집단 간의 평균차이를 보기위해 t검정과 χ^2검정을 SAS 9.2를 이용하여 분석하였다.

표 1 임상변수의 특성

변수	환자군(n=503) 빈도(%)	대조군(n=503) 빈도(%)	유의확률
당뇨	159(31.61)	7(1.39)	<.0001*
고혈압	286(56.86)	15(2.98)	<.0001*
흡연	353(70.18)	332(66.00)	0.1555
	평균±표준편차	평균±표준편차	
BMI	25.40 ± 2.94	24.24 ± 2.83	<.0001*
Tchol	190.28 ± 47.34	197.16 ± 34.71	0.0088*
HDL	41.45 ± 10.44	50.60 ± 14.11	<.0001*
TG	165.64 ± 122.27	139.26 ± 76.18	<.0001*
LDL	120.86 ± 42.09	118.80 ± 33.50	0.3909
BUN	15.59 ± 12.30	14.49 ± 3.87	0.0562
Creatinine	1.03 ± 0.45	0.86 ± 0.19	<.0001*
Glucose	133.32 ± 62.90	90.72 ± 18.26	<.0001*
Insulin	11.09 ± 9.54	7.66 ± 3.30	<.0001*
hsCRP	10.60 ± 30.12	1.19 ± 3.54	<.0001*
Adiponectin	5.20 ± 3.70	5.66 ± 2.94	0.0299*
Lp-PLA2	33.88 ± 12.86	32.89 ± 11.56	0.2014
IL6	9.62 ± 44.14	3.58 ± 8.89	0.0028*
RANTES	29758.23 ± 27075.56	30363.99 ± 23180.04	0.7037

표 2 유전자자료의 특성

SNP (rs number)		환자군(n=503) 빈도(%)	대조군(n=503) 빈도(%)	유의확률
rs10946398	AA	124(24.7)	136(27.3)	0.6217
	CA	252(50.2)	246(49.3)	
	CC	126(25.1)	117(23.5)	
rs1501299	GG	241(48.2)	218(44.0)	0.2315
	GT	205(41.0)	230(46.4)	
	TT	54(10.8)	48(9.7)	
rs16874954	GG	377(75.1)	384(77.0)	0.4761
	GT	113(22.5)	108(21.6)	
	TT	12(2.4)	7(1.4)	
rs17465637	AA	86(17.1)	97(19.4)	0.1237
	CA	237(47.2)	254(50.9)	
	CC	179(35.7)	148(29.7)	
rs1871388	GG	82(16.4)	102(20.5)	0.1867
	GT	222(44.3)	220(44.2)	
	TT	197(39.3)	176(35.3)	
rs2241766	GG	39(7.8)	34(6.8)	0.5027
	GT	213(42.5)	199(39.9)	
	TT	249(49.7)	266(53.3)	
rs4402960	AA	36(7.2)	46(9.2)	0.4938
	CA	218(43.5)	210(42.2)	
	CC	247(49.3)	242(48.6)	
rs8050136	AA	10(2.0)	7(1.4)	0.3762
	CA	120(24.1)	104(21.0)	
	CC	369(74.0)	384(77.6)	
rs11048979	CC	17(3.4)	41(8.2)	0.0019*
	TC	175(35.1)	188(37.6)	
	TT	306(61.5)	271(54.2)	
rs12713259	CC	39(7.8)	27(5.4)	0.0557
	TC	209(41.6)	186(37.1)	
	TT	254(50.6)	289(57.6)	
rs13266634	CC	203(40.4)	183(36.6)	0.3628
	TC	226(45.0)	247(49.4)	
	TT	73(14.5)	70(14.0)	
rs1502017	CC	257(51.2)	279(55.6)	0.2992
	TC	203(40.4)	190(37.9)	
	TT	42(8.4)	33(6.6)	

표 2 유전자자료의 특성(계속)

SNP (rs number)		환자군(n=503) 빈도(%)	대조군(n=503) 빈도(%)	유의확률
rs16944	CC	116(23.2)	106(21.3)	0.6735
	TC	270(54.0)	268(53.9)	
	TT	114(22.8)	123(24.8)	
rs1801133	CC	148(29.5)	156(31.1)	0.6941
	TC	260(51.9)	247(49.2)	
	TT	93(18.6)	99(19.7)	
rs2107538	AA	72(14.4)	77(15.4)	0.8929
	AG	247(49.3)	244(48.9)	
	GG	182(36.3)	178(35.7)	
rs2569190	CC	78(15.7)	73(14.8)	0.9225
	TC	241(48.4)	242(48.9)	
	TT	179(35.9)	180(36.4)	
rs4149263	CC	9(1.8)	13(2.6)	0.3142
	TC	138(27.5)	153(30.8)	
	TT	355(70.7)	331(66.6)	
rs501120	CC	61(12.2)	87(17.4)	0.0596
	TC	239(47.6)	231(46.1)	
	TT	202(40.2)	183(36.5)	
rs5015480	CC	27(5.4)	18(3.6)	0.3719
	TC	147(29.4)	147(29.3)	
	TT	326(65.2)	337(67.1)	
rs564398	CC	4(0.8)	8(1.6)	0.4035
	TC	114(22.8)	105(20.9)	
	TT	383(76.5)	389(77.5)	
rs1051931	CC	339(67.5)	365(73.0)	0.1316
	CT	150(29.9)	127(25.4)	
	TT	13(2.6)	8(1.6)	
rs155948	AA	370(73.9)	396(79.0)	0.0024*
	GA	119(23.8)	104(20.8)	
	GG	12(2.4)	1(0.2)	
rs2230806	AA	96(19.2)	114(22.9)	0.3236
	GA	262(52.4)	243(48.7)	
	GG	142(28.4)	142(28.5)	
rs5215	AA	179(35.7)	165(33.1)	0.2453
	GA	224(44.7)	249(49.9)	
	GG	98(19.6)	85(17.0)	

표 2 유전자자료의 특성(계속)

SNP (rs number)		환자군(n=503) 빈도(%)	대조군(n=503) 빈도(%)	유의확률
rs909253	AA	167(33.3)	151(30.1)	0.5595
	GA	246(49.0)	255(50.9)	
	GG	89(17.7)	95(19.0)	
rs1143623	CC	85(16.9)	89(17.7)	0.9159
	GC	255(50.8)	249(49.6)	
	GG	162(32.3)	164(32.7)	
rs1800796	CC	285(57.1)	278(55.6)	0.6909
	CG	189(37.9)	191(38.2)	
	GG	25(5.0)	31(6.2)	
rs4848306	CC	134(27.0)	155(31.1)	0.3619
	TC	264(53.1)	250(50.1)	
	TT	99(19.9)	94(18.8)	
rs1333049	CC	150(30.0)	114(22.8)	0.0116*
	CG	250(50.0)	257(51.3)	
	GG	100(20.0)	130(26.0)	
rs13290387	CC	29(5.8)	24(4.8)	0.2376
	CG	145(28.9)	169(33.7)	
	GG	327(65.3)	308(61.5)	
rs4341	CC	193(38.6)	186(37.2)	0.6818
	GC	231(46.2)	228(45.6)	
	GG	76(15.2)	86(17.2)	
rs2070600	AA	6(1.2)	11(2.21)	0.4731
	GA	131(26.3)	130(26.1)	
	GG	361(72.5)	357(71.7)	

3.2 분석 대상 변수 선택

고정효과로 설정할 변수는 <표1>의 결과를 이용하였다.

표 3 고정효과로 사용된 변수

고정효과
당뇨병, 고혈압, BMI, Tchol, HDL, Tg, Creatinine, Glucose, Insulin, hsCRP, Adiponectin, IL6

임의효과로 설정할 SNP을 선택하기 위해 <표2>의 결과를 이용하여 유의확률이 낮은 순서로 3개를 선정하여 조합을 이룬 것을 3SNPs군, 차상위 유의확률을 포함하여 4개를 선정하여 조합을 이룬 것을 4SNPs군, 4SNPs군 조합을 만든 후 차상위 유의확률을 포함하여 5개를 선정하여 조합을 이룬 것을 5SNPs군이라 한다.

표 4 유전형조합 구성

유전형조합	SNP (rs number)
3SNPs군	rs11048979, rs155948, rs1333049
4SNPs군	rs11048979, rs155948, rs1333049, rs12713259
5SNPs군	rs11048979, rs155948, rs1333049, rs12713259, rs501120

3.3 유전형조합만을 고려한 임의효과모형

유전적 관련성을 알아보기 위해 임의효과로 설정한 유전형조합만을 모형에 포함한 임의효과모형의 결과로 유전형조합의 추정효과인 odds와 odds의 95% 신뢰구간을 제시한 표이다. 이때, 유전형조합의 관측빈도가 5미만인 유전형조합은 제외하였다.

3SNPs군에서는 17개의 유전형조합이 관찰되었고, 4SNPs군에서는 35개의 유전형조합이, 5SNPs군에서는 55개의 유전형조합을 관찰할 수 있었다. 각 군에서의 임의효과모형의 결과 유전형조합들의 변동성의 차이를 확인할 수 있었고, 관상동맥질환에 대해 유전적 관련성이 있음을 확인할 수 있었다.

표 5 3개 SNPs의 유전형조합 임의효과모형 결과

3SNPs군*	빈도	Odds	Odds 95% 신뢰구간	
			하한	상한
CCAACC	8	1.0518	1.0376	1.0661
CCAACG	25	0.8833	0.8723	0.8944
CCAAGG	10	0.8390	0.8278	0.8504
CCGACC	6	0.9127	0.9003	0.9254
CCGACG	8	0.8743	0.8625	0.8863
TCAACC	77	1.1487	1.1366	1.1609
TCAACG	134	0.8855	0.8772	0.8939
TCAAGG	72	0.8486	0.8395	0.8577
TCGACC	21	1.0303	1.0173	1.0435
TCGACG	40	1.0157	1.0038	1.0278
TCGAGG	12	1.0516	1.0377	1.0656
TTAACC	109	1.1307	1.1196	1.1419
TTAACG	225	1.1127	1.1033	1.1222
TTAAGG	100	0.8771	0.8684	0.8860
TTGACC	37	1.3160	1.3004	1.3318
TTGACG	64	1.0500	1.0386	1.0615
TTGAGG	31	1.1099	1.0965	1.1235

* 3SNPs군 : rs11048979, rs155948, rs1333049

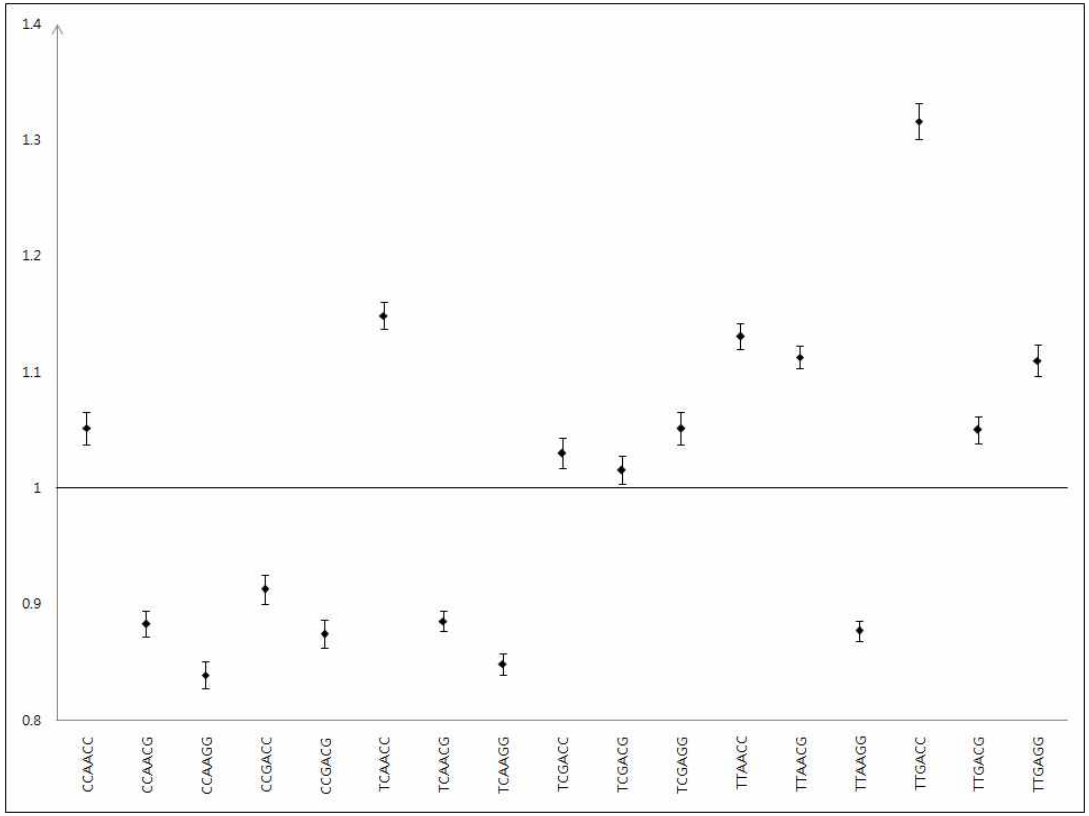

그림 1 3개 SNPs의 유전형조합 임의효과모형 Odds의 95% 신뢰구간

표 6 4개 SNPs의 유전형조합 임의효과모형 결과

4SNPs군*	빈도	Odds	Odds 95% 신뢰구간	
			하한	상한
CCAACCTT	5	1.0036	0.9983	1.0090
CCAACGTC	10	0.9937	0.9884	0.9990
CCAACGTT	12	0.9938	0.9885	0.9991
CCAAGGTT	6	0.9801	0.9748	0.9853
CCGACGTT	6	0.9801	0.9748	0.9853
TCAACCCC	6	1.0071	1.0017	1.0125
TCAACCTC	25	1.0312	1.0258	1.0367
TCAACCTT	46	0.9956	0.9904	1.0007
TCAACGCC	8	0.9936	0.9883	0.9989
TCAACGTC	53	0.9742	0.9692	0.9792
TCAACGTT	73	0.9877	0.9827	0.9928
TCAAGGCC	5	0.9968	0.9915	1.0022
TCAAGGTC	25	0.9912	0.9860	0.9965
TCAAGGTT	42	0.9702	0.9651	0.9752
TCGACCTC	10	1.0072	1.0019	1.0126
TCGACCTT	11	0.9971	0.9918	1.0024
TCGACGTC	18	0.9875	0.9823	0.9928
TCGACGTT	21	1.0109	1.0056	1.0163
TCGAGGTT	8	1.0004	0.9950	1.0057
TTAACCTC	53	1.0181	1.0129	1.0234
TTAACCTT	52	1.0085	1.0033	1.0137
TTAACGCC	16	1.0210	1.0156	1.0264
TTAACGTC	93	1.0369	1.0317	1.0421
TTAACGTT	116	0.9873	0.9824	0.9922
TTAAGGCC	7	0.9902	0.9849	0.9955
TTAAGGTC	36	1.0080	1.0028	1.0133
TTAAGGTT	57	0.9624	0.9574	0.9673
TTGACCTC	14	1.0074	1.0020	1.0127
TTGACCTT	20	1.0347	1.0292	1.0402
TTGACGCC	5	1.0174	1.0120	1.0229
TTGACGTC	18	0.9941	0.9889	0.9994
TTGACGTT	41	0.9985	0.9933	1.0037
TTGAGGCC	5	1.0036	0.9983	1.0090
TTGAGGTC	13	1.0107	1.0054	1.0161
TTGAGGTT	13	1.0039	0.9986	1.0093

* 4SNPs군 : rs11048979, rs155948, rs1333049, rs12713259

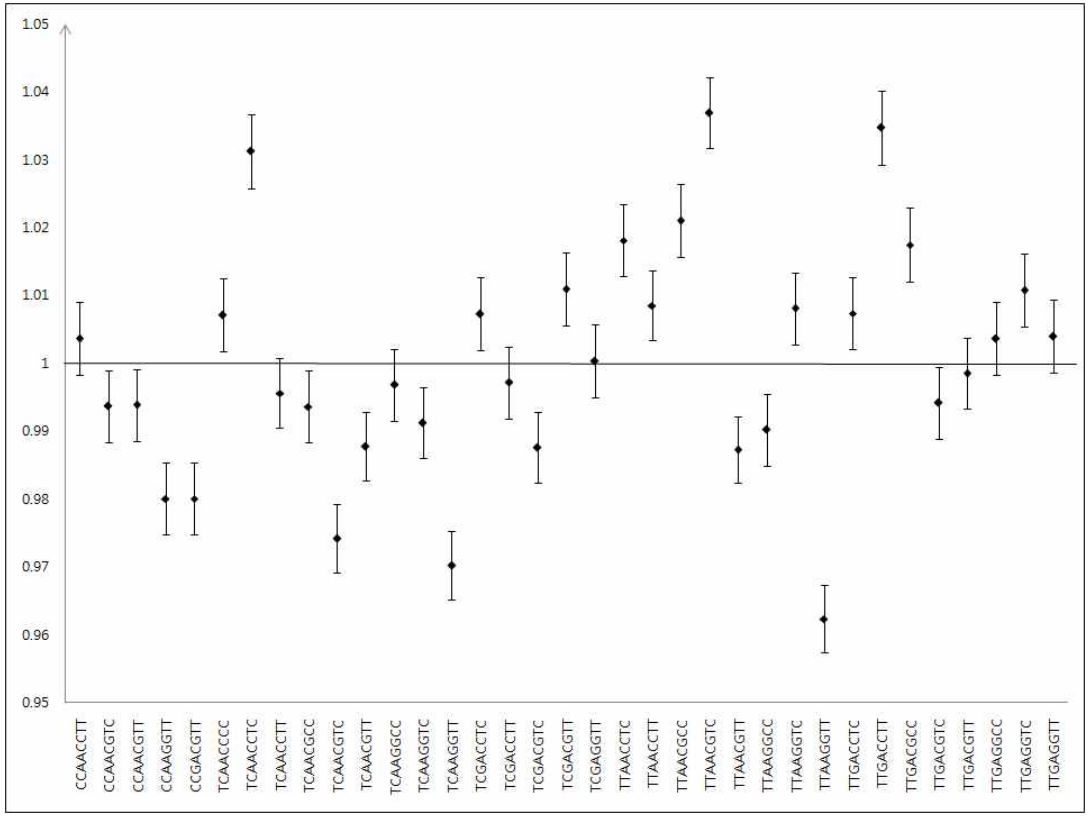

그림 2 4개 SNPs의 유전형조합 임의효과모형 Odds의 95% 신뢰구간

표 7 5개 SNPs의 유전형조합 임의효과모형 결과

5SNPs군*	빈도	Odds	Odds 95% 신뢰구간	
			하한	상한
CCAACGTCTT	5	0.9799	0.9662	0.9939
TCAACCTCCC	5	0.9799	0.9662	0.9939
TCAACCTCTC	12	1.1650	1.1491	1.1811
TCAACCTCTT	8	1.0406	1.0262	1.0553
TCAACCTTCC	7	0.9420	0.9288	0.9553
TCAACCTTTC	17	0.9473	0.9347	0.9601
TCAACCTTTT	22	1.0726	1.0586	1.0868
TCAACGTCCC	7	0.9420	0.9288	0.9553
TCAACGTCTC	21	0.9831	0.9703	0.9962
TCAACGTCTT	25	0.9193	0.9075	0.9313
TCAACGTTCC	11	0.9086	0.8962	0.9212
TCAACGTTTC	30	1.1032	1.0893	1.1174
TCAACGTTTT	32	0.9091	0.8977	0.9207
TCAAGGTCTC	12	1.0003	0.9867	1.0141
TCAAGGTCTT	10	1.0003	0.9865	1.0142
TCAAGGTTTC	23	0.9179	0.9060	0.9300
TCAAGGTTTT	16	0.9296	0.9172	0.9422
TCGACCTCTC	5	1.0208	1.0064	1.0353
TCGACCTTTC	6	1.0414	1.0268	1.0562
TCGACGTCCC	5	0.9408	0.9275	0.9542
TCGACGTCTC	6	0.9225	0.9096	0.9356
TCGACGTCTT	7	1.0620	1.0472	1.0770
TCGACGTTCC	5	0.9799	0.9662	0.9939
TCGACGTTTC	5	1.0633	1.0483	1.0785
TCGACGTTTT	11	1.0197	1.0058	1.0339
TCGAGGTTTT	6	0.9606	0.9471	0.9742
TTAACCTCCC	11	0.9812	0.9678	0.9949
TTAACCTCTC	25	0.9838	0.9711	0.9966
TTAACCTCTT	17	1.1362	1.1211	1.1515
TTAACCTTCC	8	0.9614	0.9481	0.9749
TTAACCTTTC	28	0.9062	0.8946	0.9179
TTAACCTTTT	16	1.2019	1.1858	1.2182
TTAACGCCTC	7	1.0204	1.0062	1.0348
TTAACGCCTT	5	1.0208	1.0064	1.0353
TTAACGTCCC	12	0.9629	0.9498	0.9762
TTAACGTCTC	48	1.1868	1.1727	1.2011
TTAACGTCTT	33	1.0167	1.0040	1.0296
TTAACGTTCC	21	0.9165	0.9045	0.9287

표 7 5개 SNPs의 유전형조합 임의효과모형 결과(계속)

5SNPs군*	빈도	Odds	Odds 95% 신뢰구간	
			하한	상한
TTAACGTTTC	62	0.9029	0.8926	0.9134
TTAACGTTTT	33	1.1184	1.1044	1.1326
TTAAGGTCCC	5	0.9799	0.9662	0.9939
TTAAGGTCTC	14	1.0385	1.0245	1.0527
TTAAGGTCTT	16	1.0378	1.0239	1.0519
TTAAGGTTTC	30	0.9687	0.9564	0.9811
TTAAGGTTTT	25	0.8590	0.8479	0.8702
TTGACCTCTC	11	1.0197	1.0058	1.0339
TTGACCTTTC	10	1.0003	0.9865	1.0142
TTGACCTTTT	9	1.1933	1.1768	1.2100
TTGACGTCTC	10	1.0003	0.9865	1.0142
TTGACGTCTT	7	0.9804	0.9667	0.9943
TTGACGTTTC	16	0.9644	0.9515	0.9774
TTGACGTTTT	21	1.0182	1.0049	1.0317
TTGAGGTCTC	5	1.0208	1.0064	1.0353
TTGAGGTCTT	5	1.0633	1.0483	1.0785
TTGAGGTTTT	7	0.9420	0.9288	0.9553

* 5SNPs군 : rs11048979, rs155948, rs1333049, rs12713259, rs501120

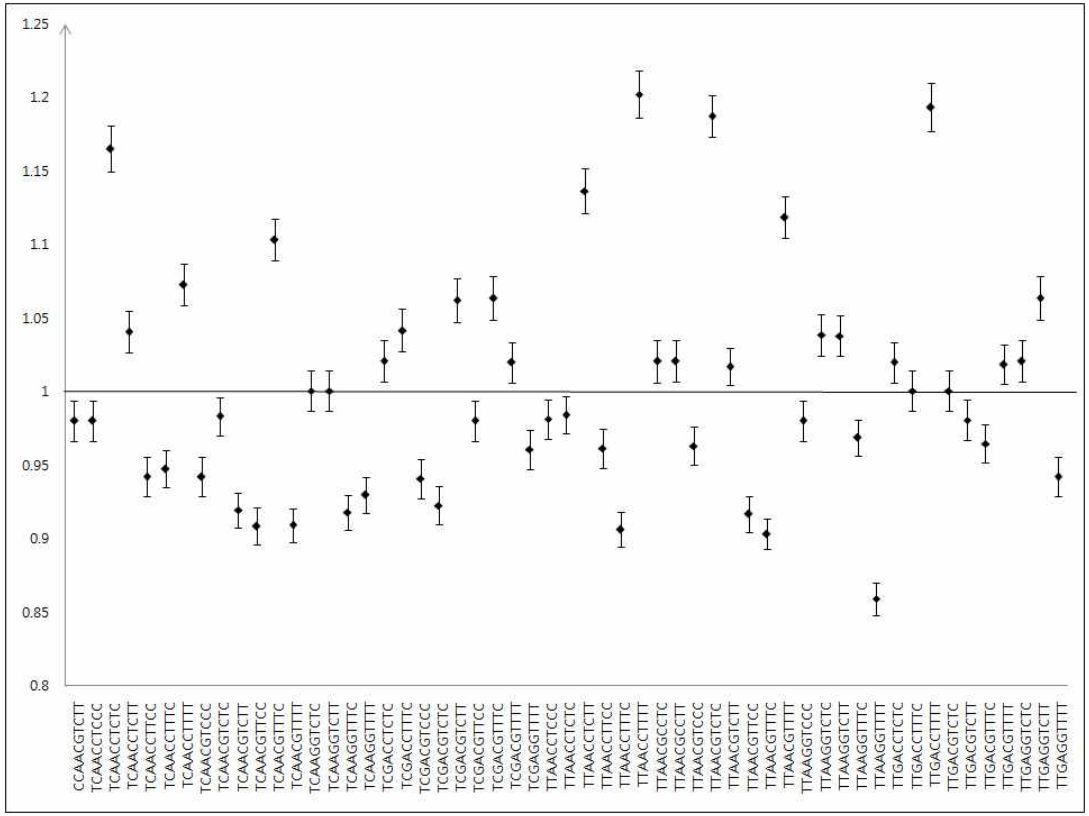

그림 3 5개 SNPs의 유전형조합 임의효과모형 Odds의 95% 신뢰구간

3.4 고정효과를 포함하는 유전형조합을 고려한 혼합효과모형

앞선 관상동맥질환에 대한 유전적 관련성을 알아보기 위한 임의효과모형에서 관련성이 있음을 확인하고, 임의효과와 <표3>에서 선택한 고정효과를 모형에 포함하여 혼합효과모형을 설정한 후 유전형조합의 추정효과인 odds와 odds의 95% 신뢰구간을 제시한 표이다. 이때, 유전형조합의 관측빈도가 5미만인 유전형조합은 제외하였다.

3SNPs군에서는 17개의 유전형조합이 관찰되었고, 4SNPs군에서는 35개의 유전형조합이, 5SNPs군에서는 55개의 유전형조합을 관찰할 수 있었다. 각 군에서의 혼합효과모형의 결과에서 관상동맥질환에 대해 유전적 관련성이 있음을 확인할 수 있었다.

표 8 3개 SNPs의 유전형조합 혼합효과모형 결과

3SNPs군*	빈도	Odds	Odds 95% 신뢰구간	
			하한	상한
CCAACC	8	1.0157	0.9996	1.0320
CCAACG	25	0.9615	0.9473	0.9760
CCAAGG	10	0.8949	0.8809	0.9092
CCGACC	6	0.9977	0.9820	1.0137
CCGACG	8	0.8746	0.8607	0.8887
TCAACC	77	1.4535	1.4339	1.4735
TCAACG	134	0.8812	0.8703	0.8922
TCAAGG	72	0.9992	0.9859	1.0127
TCGACC	21	1.0226	1.0071	1.0383
TCGACG	40	1.1848	1.1679	1.2020
TCGAGG	12	0.9246	0.9100	0.9394
TTAACC	109	1.0918	1.0777	1.1060
TTAACG	225	0.8103	0.8008	0.8199
TTAAGG	100	1.0297	1.0166	1.0430
TTGACC	37	1.2308	1.2129	1.2490
TTGACG	64	0.8895	0.8772	0.9019
TTGAGG	31	0.9171	0.9032	0.9312

* 3SNPs군 : rs11048979, rs155948, rs1333049

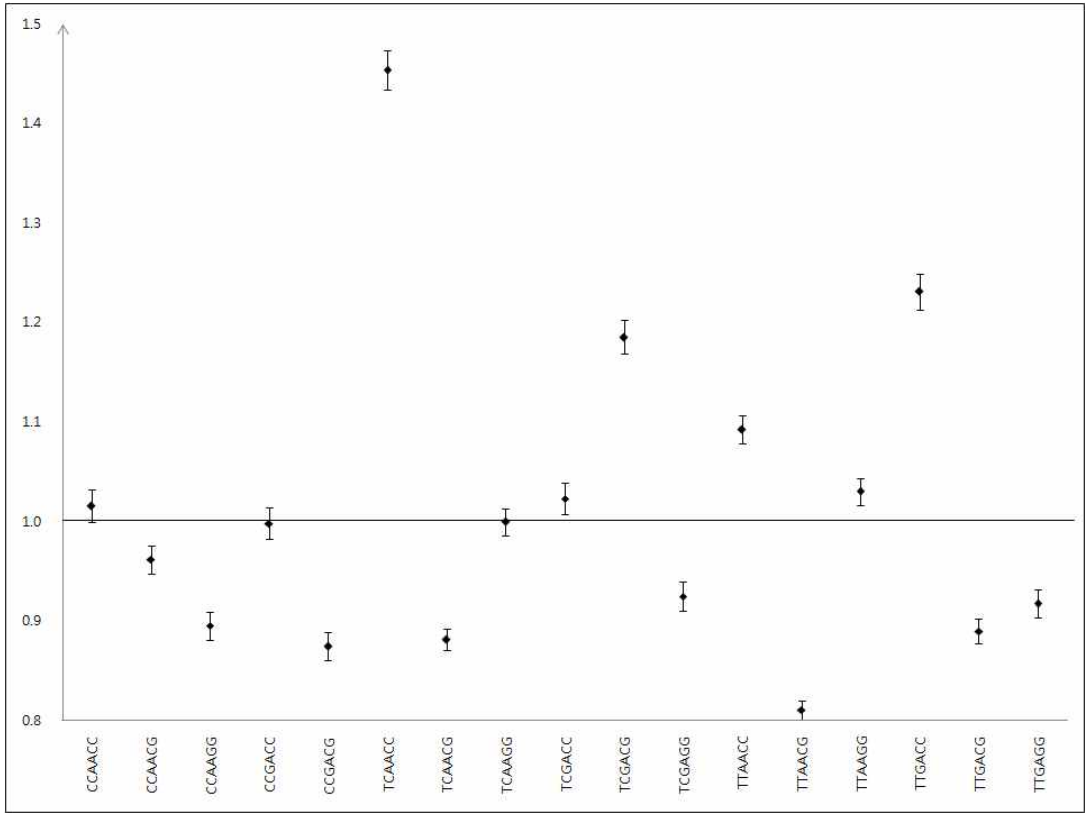

그림 4 3개 SNPs의 유전형조합 혼합효과모형 Odds의 95% 신뢰구간

표 9 4개 SNPs의 유전형조합 혼합효과모형 결과

4SNPs군*	빈도	Odds	Odds 95% 신뢰구간	
			하한	상한
CCAACCTT	5	0.9872	0.9618	1.0133
CCAACGTC	10	0.9269	0.9043	0.9502
CCAACGTT	12	1.0815	1.0559	1.1076
CCAAGGTT	6	0.8192	0.7989	0.8401
CCGACGTT	6	0.8776	0.8554	0.9003
TCAACCCC	6	1.0508	1.0237	1.0785
TCAACCTC	25	1.6351	1.5974	1.6737
TCAACCTT	46	1.3731	1.3459	1.4009
TCAACGCC	8	0.8757	0.8541	0.8979
TCAACGTC	53	0.7148	0.7006	0.7292
TCAACGTT	73	1.0299	1.0110	1.0493
TCAAGGCC	5	1.0141	0.9888	1.0399
TCAAGGTC	25	1.1114	1.0870	1.1363
TCAAGGTT	42	0.8843	0.8664	0.9025
TCGACCTC	10	1.2154	1.1857	1.2459
TCGACCTT	11	0.8611	0.8402	0.8824
TCGACGTC	18	0.8228	0.8036	0.8423
TCGACGTT	21	1.6797	1.6427	1.7175
TCGAGGTT	8	0.8235	0.8023	0.8453
TTAACCTC	53	0.8191	0.8025	0.8361
TTAACCTT	52	1.3278	1.3018	1.3543
TTAACGCC	16	1.3690	1.3363	1.4026
TTAACGTC	93	0.9131	0.8965	0.9300
TTAACGTT	116	0.5769	0.5669	0.5872
TTAAGGCC	7	1.0352	1.0100	1.0609
TTAAGGTC	36	1.1838	1.1588	1.2093
TTAAGGTT	57	0.8816	0.8648	0.8987
TTGACCTC	14	1.0936	1.0677	1.1202
TTGACCTT	20	1.3237	1.2930	1.3550
TTGACGCC	5	1.0256	0.9989	1.0530
TTGACGTC	18	1.0950	1.0707	1.1198
TTGACGTT	41	0.6950	0.6801	0.7101
TTGAGGCC	5	1.0131	0.9865	1.0404
TTGAGGTC	13	0.8922	0.8701	0.9147
TTGAGGTT	13	0.8995	0.8777	0.9219

* 4SNPs군 : rs11048979, rs155948, rs1333049, rs12713259

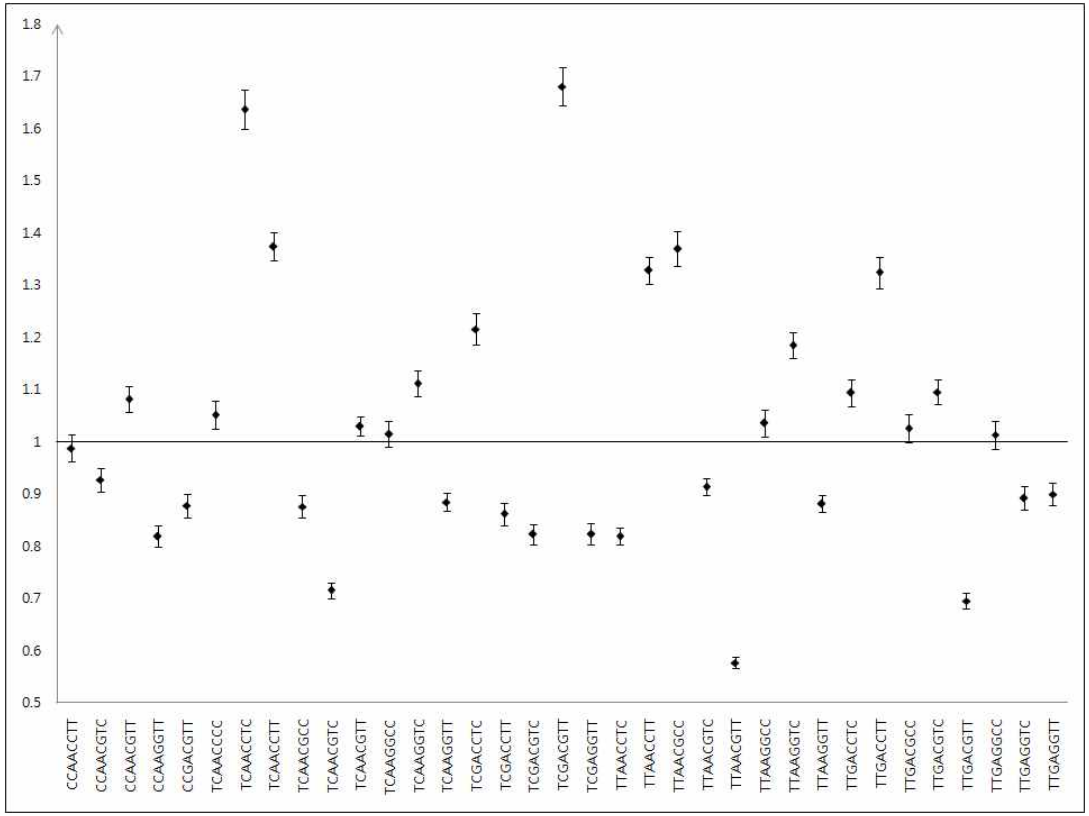

그림 5 4개 SNPs의 유전형조합 혼합효과모형 Odds의 95% 신뢰구간

표 10 5개 SNPs의 유전형조합 혼합효과모형 결과

5SNPs군*	빈도	Odds	Odds 95% 신뢰구간	
			하한	상한
CCAACGTCTT	5	0.8035	0.7704	0.8379
TCAACCTCCC	5	0.9183	0.8803	0.9580
TCAACCTCTC	12	2.3373	2.2516	2.4263
TCAACCTCTT	8	1.1833	1.1335	1.2353
TCAACCTTCC	7	0.7804	0.7492	0.8130
TCAACCTTTC	17	1.1763	1.1353	1.2188
TCAACCTTTT	22	1.8606	1.8026	1.9204
TCAACGTCCC	7	0.5444	0.5233	0.5664
TCAACGTCTC	21	1.1403	1.1019	1.1801
TCAACGTCTT	25	0.6902	0.6681	0.7129
TCAACGTTCC	11	0.7925	0.7631	0.8232
TCAACGTTTC	30	1.2453	1.2075	1.2844
TCAACGTTTT	32	0.9924	0.9617	1.0241
TCAAGGTCTC	12	1.0571	1.0186	1.0971
TCAAGGTCTT	10	1.2502	1.2041	1.2980
TCAAGGTTTC	23	0.8233	0.7967	0.8508
TCAAGGTTTT	16	0.8697	0.8395	0.9010
TCGACCTCTC	5	1.2537	1.2034	1.3061
TCGACCTTTC	6	1.0496	1.0071	1.0939
TCGACGTCCC	5	0.8197	0.7854	0.8555
TCGACGTCTC	6	0.6471	0.6212	0.6740
TCGACGTCTT	7	1.2294	1.1812	1.2795
TCGACGTTCC	5	1.2208	1.1734	1.2701
TCGACGTTTC	5	1.7291	1.6565	1.8049
TCGACGTTTT	11	1.4926	1.4398	1.5472
TCGAGGTTTT	6	0.6710	0.6420	0.7013
TTAACCTCCC	11	0.7113	0.6839	0.7397
TTAACCTCTC	25	0.8082	0.7817	0.8355
TTAACCTCTT	17	1.1283	1.0879	1.1702
TTAACCTTCC	8	0.7990	0.7681	0.8312
TTAACCTTTC	28	1.1936	1.1556	1.2329
TTAACCTTTT	16	1.8508	1.7875	1.9163
TTAACGCCTC	7	1.4813	1.4232	1.5417
TTAACGCCTT	5	1.0941	1.0488	1.1413
TTAACGTCCC	12	0.8492	0.8169	0.8827
TTAACGTCTC	48	1.1090	1.0779	1.1411
TTAACGTCTT	33	0.7725	0.7474	0.7985
TTAACGTTCC	21	0.6878	0.6642	0.7122

표 10 5개 SNPs의 유전형조합 혼합효과모형 결과 (계속)

5SNPs군*	빈도	Odds	Odds 95% 신뢰구간	
			하한	상한
TTAACGTTTC	62	0.4178	0.4065	0.4294
TTAACGTTTT	33	1.0986	1.0637	1.1346
TTAAGGTCCC	5	0.7410	0.7112	0.7721
TTAAGGTCTC	14	1.3678	1.3201	1.4173
TTAAGGTCTT	16	1.3965	1.3465	1.4485
TTAAGGTTTC	30	1.0901	1.0563	1.1249
TTAAGGTTTT	25	0.7525	0.7296	0.7762
TTGACCTCTC	11	1.2451	1.1982	1.2938
TTGACCTTTC	10	1.0631	1.0236	1.1041
TTGACCTTTT	9	1.6750	1.6086	1.7441
TTGACGTCTC	10	0.9603	0.9251	0.9968
TTGACGTCTT	7	1.2822	1.2353	1.3309
TTGACGTTTC	16	0.8815	0.8503	0.9139
TTGACGTTTT	21	0.5724	0.5527	0.5928
TTGAGGTCTC	5	0.9228	0.8839	0.9634
TTGAGGTCTT	5	0.9351	0.8958	0.9763
TTGAGGTTTT	7	0.6773	0.6508	0.7049

* 5SNPs군 : rs11048979, rs155948, rs1333049, rs12713259, rs501120

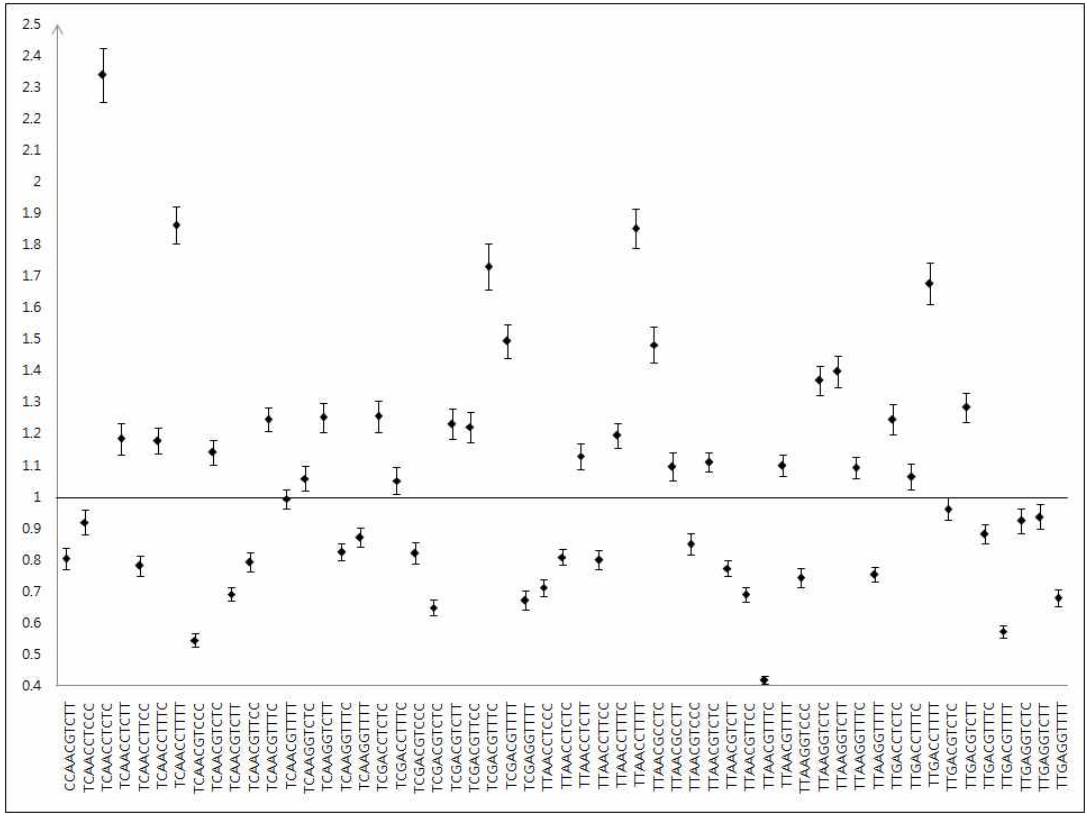

그림 6 5개 SNPs의 유전형조합 혼합효과모형 Odds의 95% 신뢰구간

3.5 로지스틱 회귀분석, 임의효과모형, 혼합효과모형의 비교

임의효과모형과 혼합효과모형에 대응하도록 가변수를 만들어 로지스틱 회귀분석에 이용하였다. 각 유전형에서는 가장 높은 빈도를 가진 $RS11048979_{TT}$, $RS155948_{AC}$, $RS1333049_{CG}$, $RS12713259_{TT}$, $RS501120_{CT}$를 참고유전형으로 각각의 가변수들의 효과를 파악하기 위해 로지스틱 회귀분석을 실시한 후 임의효과모형과 비교, 혼합효과모형과 비교하였다. 임의효과모형을 적용한 경우 예측력이 로지스틱의 회귀분석의 결과보다 3SNPs군의 경우 0.005, 4SNPs군의 경우 0.013, 5SNPs군의 경우 0.045 만큼 높은 예측력을 보였다. 혼합효과모형을 적용한 경우 예측력이 로지스틱의 회귀분석의 결과보다 3SNPs군의 경우에는 차이가 없었지만 4SNPs군에서는 0.004, 5SNPs군에서는 0.008만큼 높은 예측력을 보였다.

표 11 임의효과모형과 로지스틱 회귀분석의 AUC 비교

Type		3SNPs군[1]	4SNPs군[2]	5SNPs군[3]
임의효과모형	AUC ±S.E	0.581±0.018	0.589±0.018	0.633±0.019
로지스틱 회귀분석	AUC ±S.E	0.576±0.018	0.576±0.018	0.588±0.020

표 12 혼합효과모형과 로지스틱 회귀분석의 AUC 비교

Type		3SNPs군[1]	4SNPs군[2]	5SNPs군[3]
혼합효과모형	AUC ±S.E	0.931±0.008	0.936±0.008	0.944±0.008
로지스틱 회귀분석	AUC ±S.E	0.931±0.028	0.932±0.008	0.936±0.009

1) 3SNPs군 : rs11048979, rs155948, rs1333049

2) 4SNPs군 : rs11048979, rs155948, rs1333049, rs12713259

3) 5SNPs군 : rs11048979, rs155948, rs1333049, rs12713259, rs501120

제 4 장 결론 및 고찰

본 연구의 결과 모형을 통해 기존 로지스틱 회귀분석의 방법에서는 고려할 수 없었던 고차원 유전적 교호작용을 고려해 3SNPs군, 4SNPs군, 5SNPs군에서 분석하여 각 유전형 조합이 관상동맥에 미치는 영향이 다름을 파악해 유전적 관련성이 있다는 해석을 가능케 했다. 또한 기존에 관상동맥질환에 위험인자로 알려진 변수들을 교정효과로 모형에 추가시켜 혼합효과모형을 통해 3SNPs군, 4SNPs군, 5SNPs군에서 분석하여 역시 각 유전형 조합이 관상동맥질환에 미치는 영향이 다름을 파악해 유전적 관련성이 있음을 확인 할 수 있었다.

기존의 로지스틱 회귀분석의 방법에서는 가변수를 이용한 개별 SNP의 효과에 대한 해석만이 가능하였지만 임의효과가 포함됨으로서 개별 SNP의 효과가 아닌 유전형 조합에 따른 효과를 반영함으로써 유전적으로 개별특성을 살린 분석을 할 수 있었다. 또한, 로지스틱 회귀분석에는 가변수처리를 통한 분석시 관심 SNP가 증가함에 따라 조합의 수가 3^{SNPs}만큼씩 증가하여 주어진 자료 내에서 분석이 불가능한 경우가 발생하였지만 혼합효과모형의 이용으로 주어진 자료 안에서 분석이 가능한 장점이 있다.

부가적으로 기존의 분석방법인 로지스틱 회귀분석의 예측력을 임의효과모형과 비교해본결과 3SNPs군, 4SNPs군, 5SNPs군에서 근소하게 임의효과모형에서 AUC값이 높았던 것을 확인할 수 있었다. 그리고, 혼합효과모형의 예측력을 로지스틱 회귀분석의 예측력과 비교해본 결과 4SNPs군, 5SNPs군에서 혼합효과모형의 AUC값이 높았던 것을 확인할 수 있었다.

본 연구와 같이 혼합모형을 이용하면 여러 SNP에 의한 영향 및 SNP들과 환경적 요인과의 교호작용을 분석하면서 동시에 잠재적으로 영향을 줄 수 있는 혼란변수들을 보정할 수 있다. 그러나 혼합모형을 이용할 경우 관찰되지 않은 유전형조합에 대해서는 그 효과의 평가가 어렵다는 제한점이 존재한다.

참 고 문 헌

박용규, 송혜향. 반복측정과 교차계획자료의 분석법. *자유아카데미* 1998

명성민. HME 모형을 이용한 시간에 따른 반복측정 microarray 자료에 대한 분석기법 연구. 2006

송기준. 반복측정자료 분석을 위한 혼합모형의 임의 효과에 관한 연구. 2000

송기준, 박찬미, 임길섭, 장양수, 김동기. 혼합모형을 이용한 혈중 지질농도의 유전적 관련성 분석. *대한심장학회지* 2006:36:229-35

이준영. 일반화된 선형 혼합 모형 : 선형 혼합 모형과 일반화된 선형 모형의 연결. *응용통계연구* 1999:14:27-40

이준영. 일반화된 선형 혼합 모형에 관한 최근의 연구 동향. *응용통계연구* 2000:13:2:541-62

Breiman, L., Friedman, J., Olshen, R. and Stone, C,. Classification and Regression Trees, *Wadsworth* 1984

Brown, H. and Prescott. R. Applied Mixed Models in Medicine. *Wiley* 2006

Course Note. Statistical Analysis with the GLIMMIX Procedure. *SAS* 2007

Galwey, N. W. Introduction to Mixed Modelling. *John Wiley&Sons* 2006

Jiang. J. Linear and Generalized Linear Mixed Models and Their Applications. *Springer* 2007

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D. and Schabenberger, O. SAS for Mixed Models Second Edition. *SAS* 2006

McLachlan, G. J. and Krishnan T. The EM Algorithm and Extension. Wiley 1996

Molenberghs, G. and Verbke, G., Models for Discrete Longitudinal Data. *Springer* 2005

Shoukri, M. M. and Chaudhary, M. A. Analysis of Correlated Data with SAS and R. *Champman & Hall* 2007

West, B. T., Welch, K. B. and Galecki, A. T., Linear Mixed Models : A practical guide using statistical software. *Chapman & Hall* 2007

Breslow, N. E. and Clayton, D. G. Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* 1993:88:9-25

Foulkes, A. S., Reilly, M., Zhou, L., Wolfe, M. and Rader, D. J. Mixed modelling to characterize genotype-phenotype associations, *Statistical in Medicine* 2005:24:775-89

Harville, D. A. Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association* 1977:72:320-40

Hoh, J., Wille, A., Zee, R., Cheng, S., Reynolds, R., Lindpaintner, K. and Ott, J. Selecting SNPs in two-stage analysis of disease association data: a model-free approach, *Annals of Human Genetics* 2000;64:413 - 7

Laird, N. M. and Ware, J. H. Random effects models for longitudinal data. *Biometrics* 1982:38:963-74

Nelson, M., Kardia, S., Ferrell, R. and Sing, C. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation, *Genome Research* 2001:11:458-70

ABSTRACT

Genetic Association Analysis of CAOD Using Mixed Models

Son, Nak-Hoon
Dept. of Biostatistics and Computing
The Graduate School
Yonsei University

In genetic association studies, the complicated interaction among genes and the interaction between environmental factors and genes should be taken into consideration. However, as the number of SNPs increase, the frequently used analysis methods such as logistic regression have analytic limitation to test and interpret this complex relationships.

In this thesis, we proposed using the mixed model approach to identify significant genotype groups and the gene-gene interactions. For the purpose of these analyses, we used data from 1,006 individuals(503 individuals who were patients with CAOD and other who were healthy) and 3~5 SNPs among 32 candidate SNPs examined from Cardiovascular Genome Center, Yonsei University. We defined genotype groups which are groups of individuals with same genotypes. And these observed groups were treated as random effects in a mixed model. We compared the odds of random effects model with that of mixed effect model, and compared the result of logistic regression analysis. We could analyze genotype group-specific effects through the random effects in mixed model, and it was

possible to assess the effects of SNP combination. In conclusion, the mixed model approach provided a flexible framework for identifying a significant genetic contributions that may come about through the effects of multi-locus genotypes or through an interaction between the genotype and environmental factors with the variations in diseases.

Key Word : CAOD, mixed model, genetic association