2단계 모형을 이용한 관상동맥질환 발생 위험에 미치는 유전적 기여도 예측

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

이 혜 진

2단계 모형을 이용한
관상동맥질환 발생 위험에 미치는
유전적 기여도 예측

지도 장 양 수 교수

이 논문을 석사 학위논문으로 제출함

2007년 2월 일

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
이 혜 진

이혜진의 석사 학위논문을 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

연세대학교 대학원

2007년 2월 일

감사의 글

설레는 마음으로 대학원에 진학한지 어느새 **2**년이 지났습니다. 저에게 **2**년의 시간은 많은 고민으로 힘들었던 시간들이였고 그만큼 나 자신을 더 발전시킬 수 있는 나날들이었습니다. 맨땅에 헤딩하듯 시작한 공부가 하나의 작은 결실을 맺었습니다. 공부할수록, 새롭게 헤쳐가야 할 바다가 넓기만 한 의학통계 분야인데, 아주 작은 점과도 같은 논문을 쓰면서도 허덕거리며 많은 분들의 도움을 받았습니다.

먼저 바쁘신 중에도 좋은 가르침을 주신 장양수 교수님께 진심으로 감사드립니다. 의학 통계학이라는 학문의 길을 열어 주셨고, 존경할 수 있었던 고 김동기 교수님께 감사드립니다. 의학통계에 발을 디딜 수 있게 도움을 주신 조진남, 김동건 교수님께 감사드립니다. 유전통계의 관심을 갖을 수 있었던 것은 임길섭 교수님의 가르침 이 있었기 때문입니다. 선생님의 가르침에 힘입어 관심있는 분야에 대해 논문을 쓸 수 있어 더욱 보람된 시간이었습니다.

마지막까지 인내하시며 지도해주신 송기준 선생님께 감사의 마음을 전합니다. 부족한 저에게 관심 가져 주시고 열정을 다해 가르쳐주셨던 찬미언니에게 감사의 마음을 전합니다. 가장 힘든 시기에 위로와 격려로 함께해주시고 망막함에 힘들어 할 때 자신의 일처럼 걱정해주시고 도와주신 미영언니에게 진심으로 감사드립니다. 송기준 선생님과 언니들의 세밀하고 따뜻한 관심으로 이 논문이 나올 수 있었습니다. 그 학문의 열정을 본받아 부끄럽지 않은 후배가 되겠습니다. 심혈관유전체연구소에서 근무함에 감사할 수 있었던 것은 진우언니와 은정언니가 있었기 때문입니다. 모자란 동생이지만 항상 챙겨주신 고마운 언니들에게 감사의 마음을 전합니다. 항상 긍정적인 마음으로 격려해주신 최정란 선생님께 감사드립니다. 졸업 후에도 늘 변함없이 아껴주며 많은 도움을 준 성은선배에게 진심으로 고마운 마음을 전합니다. 든든한 선배가 되어주신 명성민, 한무영, 서원열 오빠들과 신영, 은혜, 은정, 명희, 현선, 정숙 언니들에게 감사의 마음을 전합니다. 함께 공부하며 동기임에도 불구하고 늘 챙겨주신 은희언니에게도 고마운 마음을 전합니다. 심혈

관유전체연구소에 올라와 낯선 환경에 적응하느라 힘들었을 진희에게 미안한 마음과 잘 적응해준 고마움을 전합니다. 앞으로 함께 발전해 나갈 수 있길 기대합니다. 더욱 친해지고 싶었던 후배 정윤언니, 경화, 수희, 영애, 성유에게 아쉬운 마음이 앞섭니다.

삶의 버팀목이 되어주며 중보로 든든한 힘이 되어준 남주, 형구, 기웅, 창배, 정윤에게 고마운 마음과 사랑하는 마음을 전합니다. 항상 내편이 되어주는 사랑하는 영미에게도 고마운 마음을 전합니다. 나의 친구라는 것이 자랑스러운 지혜, 경연, 자량, 소연에게 너무 고맙습니다. 비록 대학원은 다르지만 같은 고민을 하며 혼자 걷고 있지 않음을 느끼며 의지할 수 있었던 은경에게도 고마움을 전하며, 지영, 아영, 성희, 지은에게도 함께 해줘 고맙다는 말을 전하고 싶습니다. 시끌벅적하지만 초등학교 때부터 함께하며 힘이 되어준 민경, 경환, 현수, 상준, 영관, 진영, 종문에게도 고마운 마음을 전합니다. 항상 웃음과 재치로 생활의 기쁨을 준 용석, 진호, 노준 정말 고맙습니다. 언제나 잘할 수 있다고 격려해준 민기에게 고마움을 전합니다. 모두들 사랑합니다. 그리고 목자 되신 박승준 목사님, 박기명 목사님께도 감사드립니다.

무조건적인 사랑으로 키워주신 부모님께 감사드립니다. 어떠한 일이든 이해해주시고 믿어주시고 인내하며 기다려주신 부모님과 사랑하는 언니와 선호 덕분에 여기까지 올 수 있었습니다. 정말 감사하고 또 감사하며 사랑합니다. 얼마 남지 않은 언니의 결혼을 진심으로 축복합니다.

항상 손녀딸을 자랑스러워 해주신 돌아가신 할아버지가 너무 보고싶습니다. 항상 존경하는 할아버지의 부끄럽지 않은 손녀딸이 되도록 노력하겠습니다. 항상 맘속에서 전하지 못한 말 “사랑합니다.“

그리고 내가 사는 이유, 나의 하나님께 감사와 찬양을 드립니다. 늘 부족한 자에게 생각지 못한 은혜로 베풀어 주신 하나님께 감사드립니다.

2007년 **2**월

이 혜 진 올림

차 례

표 차 례

그 림 차 례

국문요약

2단계 모형을 이용한 관상동맥질환 발생 위험에 미치는 유전적 기여도 예측 모형

관상동맥질환(Coronary artery disease)은 다인자성 질환(Complex disease)으로 여러 유전자 변이의 복합적으로 영향과 비만, 흡연, 음주, 지질농도 등의 이미 보고되어진 다양한 위험요인에 영향을 받는 것으로 알려져 있다.

본 논문에서는 기존의 위험인자로 보고되어진 양적형질들을 분리하여 매개변수 형질(Intermediate Phenotype, IP)로 정의한 단계적 모형 즉, 질병 발생과 관련된 위험인자들을 효과적으로 조절하기 위한 **2**단계 모형(Two-Stage Model)을 제안하였다. **2**단계 모형에서 첫 번째 단계에서는 후보 유전자와 매개변수 형질과의 연관성을 평가하기 위한 회귀분석을 수행하며, 두 번째 단계에서는 관상동맥질환에 대한 유전적 위험도를 평가하기 위한 로지스틱 회귀분석을 수행했다.

2단계 모형의 적용 결과, 공통의 매개변수 형질에 영향을 미치는 다수의 유전자들의 높은 정보력을 이용하여 질병 발생 위험에 미치는 유전적 기여도를 좀 더 구체적으로 평가할 수 있다. 또한 매개변수 형질을 통해 계산된 유전적 기여도(Genetic Risk Score)를 이용하여 유전적 효과 자체만을 고려했을 때 관상동맥질환 발생에 미치는 유전적 영향력을 확인할 수 있었다.

핵심 되는 말 : **2**단계 모형, 양적형질, 매개변수 형질, 회귀분석, 로지스틱 회귀분석, 유전적 기여도.

1장 서 론

1.1 연구배경

고혈압, 동맥경화증 등의 대부분의 심혈관계질환은 다인자성 질환(Complex disease)으로 다양한 유전자 변이가 복합적으로 질병발생과 기전에 영향을 미치는 것으로 알려져 있다. 따라서 다요인적인 유전적 위험요인을 평가하기 위한 방법을 필요로 한다.

다수의 유전자와 환경적 위험요인(Risk factor)은 심혈관계 질환과 같은 일반적인 질병에 기여할 것이다. 비록 명확한 유전적 원인에 대해서는 질병의 과거력이나 확실한 가족력에 의해 확인되어 질 수 있지만, 이러한 일반적인 모집단에 대한 질병과 관련된 유전자의 연관성에 대해서는 충분히 설명될 수 없다. 그 이유 중 하나로 질병의 유전적 경향은 하나의 단일 유전자에 의한 영향이기 보다는 다수의 유전자의 변동으로 누적된 효과에 의한 결과일 수 있기 때문이다. 또한, 질병의 유전적 경향의 변화를 주는 환경적 요인이 미치는 영향이 고려되어져야 하기 때문이다. 이렇게 질병에 관련된 유전적 효과를 밝히고자 하는 이유는 유전자가 직·간접적으로 관여하는 질병에 대한 질병유전자를 밝히는 노력으로 병의 발생을 늦추거나 증상을 완화시킬 수 있으며 질병관련 유전위험을 미리알고 질병이 나타나는 것을 미리 예방할 수도 있기 때문이다. 이러한 점을 반영하여 질병에 미치는 유전자의 영향력을 찾고자 질병과 관련된 유전자들의 연관성 분석연구가 많이 이루어져 있으며, 이때 환경적 요인이 미치는 영향을 고려하기 위한 방법으로 이들의 위험요인을 보정한 분석방법을 사용하고 있다. 하지만, 질병에 미치는 유전적 효과(Genetic effect)는 약 10%내외로 매우 작은 부분의 영향력으로 인해 상대적으로 질병에 미치는 영향력이 큰 위험인자를 보정한 방법으로 질병과 유전적 영향의 연관성을 찾기란 매우 어려운 일이다.

따라서 질환 발생예방을 위해 유전적 영향을 미리 예측하고 위험요인들을 구체적으로 조절할 수 있는 접근 방법이 요구된다.

1.2 연구목적 및 내용

다인자성 질환 발생과 관련하여 알려진 후보 유전자(Candidate gene)의 다형성이 질병발생과 관련된 독립적인 위험인자는 아닐지라도 기존의 알려진 위험요인에 의해 위험도를 높이는 것으로 보고되어져 있다. 이러한 위험인자에 대한 영향력을 충분히 갖는 사람은 질병의 위험도를 증가시킬 것이다. 본 논문에서는 질병 발생과 관련된 위험요인들을 조절하기 위한 방법으로 **2**단계 모형(Two-Stage Model)을 제안한다.

기존의 위험인자들을 보정하여 유전자와 질병의 연관성을 분석하였던 접근방법과 달리 기존의 위험요인으로 보고되어진 양적형질들을 분리하여 매개변수 형질(Intermediate Phenotype; IP)로 정의한 단계적 모형으로, 위험인자들을 효과적으로 조절하기 위한 모형이다. **2**단계 모형 적용을 통하여, 공통의 매개변수 형질에 영향을 미치는 다수의 유전자들의 높은 정보력을 이용하여 질병 발생에 미치는 유전적 기여도를 평가할 수 있다.

어떤 하나의 유전자가 다인자성 질환에 미치는 영향력의 정도는 같은 매개변수 형질에 의해 통제되는 많은 유전자에 잠재적이기 때문에 그 영향력의 정도가 작을 것이며, 실제로 다중의 매개변수 형질에 영향을 미치는 다수의 유전자가 임상적 결과에 작용할 것이다. 이러한 이유로 많은 단일 염기 다형성들 중 다중의 매개변수 형질에 대한 높은 정보력을 갖는 단일 염기 다형성들을 선별하여 **2**단계 모형에 적용한다. 따라서 다수의 유전자들 중 공통의 매개변수 형질에 잠재적인 연관성을 갖는 다수의 단일염기 다형성들은 단일 연관성을 보던 방법보다 더 높은 정보력을 갖는 유전적 영향력을 이용할 수 있으며, 질병 발생 위험에 미치는 유전적 기여도를 토대로 다인자성 질환에서 환경적 요인과 유전적 요인의 조합에

따라 질병 발생의 차이를 확인할 수 있으므로 질병 원인에 좀 더 구체적으로 접근할 수 있다.

본 논문에서는 관상동맥 질환 발생에 미치는 유전적 기여위험도를 예측하기 위한 모형으로 매개변수 형질을 고려한 **2**단계 모형을 적용하여 유전적 요인과 환경적 요인의 조합에 따라 관상동맥질환 발병 확률의 차이를 확인하고자 한다.

논문에 사용한 분석 대상은 연세대학교 심혈관계질환 유전체 연구센터(Cardiovascular Genome Center, CGC)에서 수집된 자료로 관상동맥질환 진단을 받은 환자군과 정상 대조군으로 구성된 총 **1311**명을 연구대상으로 하였다. 연구에 사용된 후보 유전자는 심혈관계질환 유전체 연구센터 자료로 9개의 유전자에 위치한 **2**3개의 후보 유전자들을 분석에 사용하였다. 심혈관계질환 유전체센터에서 수집된 관상동맥 질환 진단을 받은 대상자와 정상 대조군의 프로토콜(Protocol)을 토대로 수집된 자료를 이용하였다.

1.3 논문 구성

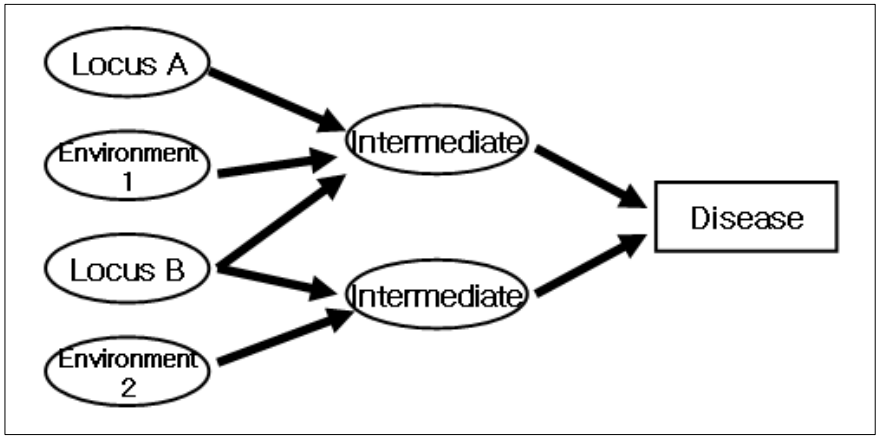
제 **1**장 서론에서는 연구의 배경에 대해 소개하고 연구 목적 및 내용에 대해 제시한다. **2**장에서는 **2**단계 모형의 이론적 배경과 각 단계의 모형에서 사용된 회귀분석 방법과 로지스틱 회귀분석 방법에 대해 설명한다. 기본적인 선형모형과 로짓 모형에서의 모수 추정 방법에 대해 간략히 설명하였다. 3장에서는 **2**단계 모형을 적용한 분석 방법에 대해 소개 한다. 질병 발생 위험에 미치는 유전적 기여도를 계산하기 위한 방법을 제안하였다. 4장에서는 분석에 사용한 심혈관계질환 유전체연구센터 자료(CGC자료)에 대해 서술하였으며 **2**단계 모형에 분석되어질 분석 대상 유전자 선택 방법과 **2**단계 모형에 적용한 결과를 제시하였다. **2**단계 모형을 이용해 질병 방생 위험에 미치는 유전적 기여도를 가지고 관상동맥 질환 발생 확률을 예측하였다. 5장에서는 연구 결과에 대한 설명과 결론을 발표 하고, 앞으로 진행되어야 할 연구에 대한 제안점에 대해 서술하였다.

2장 2단계 모형 이론

2.1 이론적 배경

2단계 모형은 단일 염기 다형성(Single nucleotide polymorphism: SNP)과 질병과의 직접적인 연관성보다 질병에 영향을 미치는 대부분의 위험 요인을 매개변수 형질로 정의하여 공통의 매개변수 형질에 영향을 받는 다수의 단일 염기 다형성들의 정보력을 이용하는 모형이다. 즉 질환에 미치는 영향은 단일 요인에 의해 설명 되어지는 변동의 일부분보다 많은 위험 인자들의 공통적인 요인인 매개변수 형질을 고려하여 설명할 수 있는 변동 부분을 향상시킬 수 있는 모형이다. 질환에 대한 대부분의 위험 인자로 알려진 요인들을 실제로 매개변수 형질이라 정의 하였다. (Green, 2004) 질환은 통합적인 매개변수 형질에 의해 영향을 받으며 공통의 매개변수 형질에 관련된 다수의 환경인자와 유전자들로 이루어진 가장 유사한 집합을 구성할 수 있다.

어떤 하나의 유전자가 다인자성 질환에 미치는 영향력의 정도는 같은 매개변수 형질에 의해 통제되는 많은 유전자에 잠재적이기 때문에 영향력의 정도가 작을 것이며, 실제로 다중의 매개변수 형질에 영향을 미치는 다수의 유전자가 질환에 작용할 것이다. 따라서 그러한 형질을 충분히 갖는 사람은 질병의 위험을 증가시킬 것이며 위험 인자의 원인이 되는 유전적, 환경적 인자와 그들의 조합은 질병 발생에 미치는 유전적 기여도에 따른 발생확률의 차이를 보일 것이다. [그림 1]은 2단계 모형의 기본적인 모형도 이다.



[그림 **1**] **2**단계 모형도

2단계 모형도는 매개변수 표현형(Phenotype)에 선형 모형을 확장시킨 i번째 유전자형(Genotype)에 노출된 j번째 환경적 영향으로 구성된 선형 함수(Linear function)모형으로

$$z_{ij} = G_i + E_j + e_{ij}$$

과 같이 나타낸다. 이때 G_i 는 표현형 발현(Phenotypic expression)에 미치는 유전적 효과를 나타내며, E_j 는 표현형 발현에 미치는 환경적 효과를 타나낸다. 유전자형 값인 G_i은 E_j보다 표현형 발현에 미치는 영향력이 매우 작기 때문에 G_i와 E_j를 동등한 독립변수로 분석하는 것은 그 유의성을 평가하기에 적절치 않다고 생각하여 **2**단계 모형에서는 유전적 영향(Genetic effects)에 미치는 부분과 환경적 영향(Environmental effects)에 효과를 나타내는 회귀계수를 각각 산출하여, **2**단계 모형의 첫 번째 단계인 매개변수 형질에 미치는 유전자의 영향력을 구한 점수에 이용한다. 다음의 식을 적용하여 얻어진 각각의 회귀계수를 적용한다.

$$z_i = G_i + e_i$$
$$z_j = E_j + e_j$$

즉 **2**단계 모형은 G_i , E_j 와 매개변수 형질과의 잠재적인 연관성을 분석하여 유전자의 영향력에 대한 점수를 구한 후, 매개변수 형질과 질환과의 연관성 분석으로 얻어진 회귀계수만큼 가중치를 주어 전체 기전(Pathway)에 미치는 유전적 영향에 대한 기여위험점수를 구한다. 즉 유전적 기여위험도는 유전자의 공유된 매개변수 형질을 토대로 생리학적 경로의 유전적 특성을 나타내는 지표로 볼 수 있다.

2.2 2단계 모형 소개

2.2.1 첫 번째 단계 : 선형 회귀모형

회귀모형은 관심있는 종속변수와 유전적, 환경적 요인의 관계를 파악하기 위한 방법이다. 종속변수인 표현형 y에 대해 a는 y의절편이고, 유전적 요인에 의해 좌우되는 효과인 b와 오차인 ϵ를 고려한 선형 회귀모형(Linear Regression Model) 식은

$$y_i = a + bG_i + \epsilon$$

이 된다. y는 종속변수 또는 반응변수라 하고 G는 독립변수 또는 설명변수라 한다. ϵ는 관계식으로 통제할 수 없는 부분으로 오차라 하며 평균이 0인 정규분포를 가정한다.

각각의 후보유전자와 형질간의 관련성 여부에 대해 귀무가설 ($H_0 : b_i = 0$)

을 검정한다. 후보유전자의 주효과(Main effect)에 대해 모형으로부터 추정된 효과인 회귀계수 b를 추정한다. 모형에 의해 예측된 y의 값에 대한 모형은 다음과 같다.

$$\hat{y_i} = a + bG_i$$

가장 적합한 a, b를 추정하기 위한 방법으로 양적형질의 유전 양식을 연구하는 유전학의 한 분야인 양적유전학(Quantitative genetics)의 관점에서 표현형 자료로부터 통계량을 쉽게 구할 수 있기 때문에 매우 유용한 특징을 지닌 잔차의 제곱합을 최소화 하는 최소제곱법(Ordinary Least Squares, OLS)을 사용하여 추정한다. 즉 최소제곱법은 회귀선에 의해 예측된 값으로부터 관찰된 y의 편차 제곱의 평균값을 최소화 시킨 값이다. n명의 개인별 표본에서 G 와 y에 대한 측정값을 얻은 후 잔차의 정의를 기억해 보면, 식은

$$\epsilon = y - \hat{y} = y - a - bG$$

$$\epsilon = (y - \overline{y}) - b(G - \overline{G}) - (a + b\overline{G} - \overline{y})$$

이 된다. 위의 식에 양변을 제곱 시키면,

$$\epsilon^2 = (y - \overline{y})^2 - 2b(y - \overline{y})(G - \overline{G}) + b^2(G - \overline{G})^2 + (a + b\overline{G} - \overline{y})^2$$

$$- 2(y - \overline{y})(a + b\overline{G} - \overline{y}) + 2b(G - \overline{G})(a + b\overline{G} - \overline{y})$$

인 식으로 나타내며, 마지막으로, ϵ^2의 평균값을 고려한 식은

$$\overline{\epsilon^2} = (\frac{n-1}{n})[\,Var(y) - 2bCov(G,y) + b^2\,Var(G)] + (a + b\overline{G} - \overline{y})^2$$

이 되고, $(G-\overline{G})$와 $(y-\overline{y})$의 평균값은 정의에 의해 '0'의 값을 갖는다. 따라서, 위와 같이 최소화된 $\overline{\epsilon^2}$를 갖는 a, b의 값은 그들이 0과 같다고 설정하고 편미분도함수(Partial derivatives)를 구하면,

$$\frac{\delta(\overline{\epsilon^2})}{\delta a} = 2(a+b\overline{G}-\overline{y}) = 0$$

$$\frac{\delta(\overline{\epsilon^2})}{\delta b} = 2[(\frac{n-1}{n})[Cov(G,y)+b\,Var(G)]+\overline{x}(a+b\overline{G}-\overline{y})] = 0$$

인 방정식으로 나타낼 수 있다. 위의 두 식을 풀면 아래의 식을 얻을 수 있다.

$$a = \overline{y}-b\overline{G}$$

$$b = \frac{Cov(G,y)}{Var(G)}$$

다음과 같이 구한 회귀계수는 독립변수의 단위를 그대로 반영한 비표준화회귀계수(Unstandardized coefficient)이다.

여러 독립변수들이 종속변수에 미치는 영향력의 크기에 대해 독립변수들의 단위와 분포를 통일 시켜주기 위해 **2**단계 모형에서는 비표준화회귀계수를 변화시킨 표준화회귀계수(Standardized coefficient)를 사용한다. (Horne et al., **2**005)

표준화 회귀계수는 회귀계수에 x의 표준편차 $SD(x_i)$를 y의 표준편차 $SD(y)$로 나누어진 값을 곱하여 다음과 같이 표현된다.

$$\widehat{b_i^s} = \hat{b}_i SD(x_i) / SD(y)$$

2.2.2 두 번째 단계 : 로지스틱 회귀모형

두 번째 단계에서는 종속변수가 이분형 변수로 일반 회귀모형에 그대로 적용하기가 매우 어려워 이 자료에 로짓 변형(Logit transformation)을 시켜 일반적인 회귀모형의 형태를 지닌 로지스틱 회귀분석(Logistic regression Analysis)을 사용한다.

질병발생여부를 나타내어 주는 종속변수로 질병 발생은 D=**1**, 질병 미발생은 D=0에 대해 probit 모형은

$$D = \beta_0 + \beta_1 x + e$$

이 되고, 전체 표본 수는 n_i, 환자군은 Y=**1**, 대조군은 Y=0 에 대한 우도 함수(Likelihood function)는

$$\prod_{i=1}^{n_1} P(x_i | Y_i = 1, D_i = 1) \prod_{i=1}^{n_0} P(x_i | Y_i = 0, D_i = 1)$$

이 된다. 위의 식에서 우도 함수의 개별 부분에 대해 베이지안 이론(Bayes Theorem)을 적용하면

$$P(x | Y, D = 1) = \frac{P(Y | x, D = 1)\, P(x | D = 1)}{P(Y | D = 1)}$$

으로 나타내며, 분자의 $P(Y | x, D = 1)$에 두 번째 베이지안 이론을 적용 하면 Y=**1** 일 때 식은

$$P(Y = 1 | x, D = 1) = \frac{P(Y = 1 | x)\, P(D = 1 | x,\ Y = 1)}{P(Y = 0 | x) P(D = 1 | x,\ Y = 0) + P(Y = 1 | x) P(D = 1 | x,\ Y = 1)}$$

으로 나타낼 수 있다. 환자군과 대조군의 선택이 독립적임을 가정할 때, 각각의 확률인 τ_1 과 τ_0은

$$\tau_1 = P(D=1|Y=1,x) = P(D=1|Y=1)$$

$$\tau_0 = P(D=1|Y=0,x) = P(D=1|Y=0)$$

와 같이 구할 수 있으며, 로지스틱 모형에 추정된 τ_1 과 τ_0을 대입하면,

$$P(Y=1|x,D=1) = \frac{\tau_1\pi(x)}{\tau_0[1-\pi(x)]+\tau_1\pi(x)} = \pi^*(x)$$

이 된다. 위의 식에서 분자와 분모에 $\tau_0[1-\pi(x)]$로 나누어 주면 로지스틱 회귀모형에서 절편 항인 $\beta_0^* = \ln(\frac{\tau_1}{\tau_0})+\beta_0$로 나타내어 진다. 표본은 독립임을 가정했기 때문에 $P(x|D=1) = P(x)$이며, 다시 정리한 식은

$$P(x|Y=1,D=1) = \frac{\pi^*(x)P(x)}{P(Y=1|D=1)}$$

로 표현된다. 또한 Y=0 일 때 식은 다음과 같다.

$$P(x|Y=0,D=1) = \frac{[1-\pi^*(x)]P(x)}{P(Y=1|D=1)}$$

우도 함수에 대한 식은

$$L^*(\beta) = \prod_{i=1}^{n} \pi^*(x_i)^{Y_i}[1-\pi^*(x_i)]^{1-Y_i}$$ 일때,

으로 표현되어진다. 즉, 위에서 구한 우도 함수를 다시 쓰면,

$$\prod_{i=1}^{n_1} P(x_i|\, Y_i = 1, D_i = 1)\prod_{i=1}^{n_0} P(x_i|\, Y_i = 0, D_i = 1) = L^*(\beta)\prod_{i=1}^{n}[\frac{P(x_i)}{P(\,Y_i|D_i = 1)}]$$

식과 같다. 최대 우도 추정은 모집단에서 얻어진 표본들이 독립이라는 가정 하에 로그 우도 함수의 최대화로부터 구해지는 추정치이므로 위의 우도 함수에 로그를 취한 로그우도 함수는

$$\ln L = \sum_{i=1}^{n} Y_i \ln[\frac{P(x_i)}{1-P(x_i)}] + \sum_{i=1}^{n} \ln[1-P(x_i)]$$

이 된다. 위의 로그 우도 함수를 최대화 시키는 회귀계수의 최대우도 추정치는 선형이 아니므로 뉴튼-랩슨(Newton-Raphson)방법이나, 피셔의 스코어링방법 (Fisher's Method of scoring)등과 같은 반복적인(iterative) 추정 방법에 의해 근사해를 구한다.

3장 2단계 모형 적용

3.1 첫 번째 단계 : SNP Analysis

선택되어진 모든 후보 유전자에서 매개변수 형질의 연관성은 결정되었고, 매개변수 형질에 개개인의 영향에 의한 각각의 유전형의 효과만큼 가중치를 부여하여 일반화한 유전적 기여위험 모형을 구현한다.

만약 다수의 단일 염기 다형성들이 단일 유전자(Single gene)나 다중 유전자(Multiple gene)에 이용된다면 물리적으로 가깝기 때문에 독립적인 단일 염기 다형성들을 모형에 고려하였다.

독립변수로서의 단일 염기 다형성과 각각의 연속형 변수로 이루어진 종속변수로서의 매개변수 형질과의 연관성을 평가하기 위해 선형 회귀분석을 수행한다. 또한 환경적 요인과 매개변수 형질과의 선형 회귀분석을 수행한다. **1**장의 연구배경에서 언급하였듯이 환경적 요인을 보정하여 분석하던 기존의 방법과 달리 각각 회귀분석을 수행한 이유는 유전적 요인이 미치는 영향에 비해 환경적 요인이 양적형질에 미치는 영향력이 상당히 크므로 유전적 요인과 환경적 요인을 동등한 독립변수로 고려할 경우 유의한 유전적 영향을 찾아내기란 매우 어려운 일이기 때문에 각각의 회귀분석을 수행하여 구한 회귀계수를 유전적 기여위험도를 구한 점수에 가중치를 주어 적용시킨다. 공통된 집합으로 구성된 단일염기 다형성들과 각각의 매개변수 형질에 대한 분석을 반복한다. 각각의 회귀분석 수행 결과 매개변수 형질의 변화를 유의하게 예측하는 독립변수들의 회귀계수를 이용한다.

이들의 선형 회귀분석으로 얻어진 비표준화회귀계수를 변형시킨 표준화회귀계수(β_{ij})만큼 가중치를 주어 매개변수 형질에 영향을 미치는 후보 유전자들의 영향력을 구한 점수(Genetic Risk Score, GRS_i)를 산출한다. i번째 매개변수 형질에 대한 Genetic Risk Score의 계산 공식은 다음과 같다. (Horne et al, **2**005)

$$GRS_i = \sum \beta_{ij} S_j$$

이때 S_j는 j번째 SNP 변수이며, β_{ij} 는 i번째 매개변수 형질과 j번째 SNP 변수 또는 환경적 요인변수에 대한 표준화회귀계수를 나타낸다.

3.2 두 번째 단계 : IP Analysis

각각의 매개변수 형질들은 종속변수인 질병에 예측자(Predictor)로서 두 번째 회귀분석을 하여 평가한다. 매개변수 형질이 질환 발생과 어떤 관련성이 있는지 알아보기 위해 로지스틱 회귀분석(Logistic Regression Analysis)을 수행한다. 두 번째 단계에서는 질환 발생에 영향을 주는 매개변수 형질에 대한 영향력을 계산하는 단계이다.

위의 분석을 통해 얻어진 결과를 토대로 전체적인 경로의 유전적 기여 위험도에 대한 수치를 나타내는 전체적인 기전에 영향을 미치는 유전적 기여도(overall pathway's Genetic Risk Score, GRS_{tot})를 구한다. 전체 경로의 Genetic Risk Score는 개개인의 i번째 매개변수 형질에 대한 Genetic Risk Score에 종속변수에 유의한 영향을 미치는 매개변수 형질들의 회귀계수 중 가장 큰 값을 갖는 회귀계수를 각각의 회귀계수에 나누어준 값만큼의 가중치를 주어 산출된다. 즉 전체 경로에 대한 Genetic Risk Score 계산 공식은 다음과 같다. (Horne et al, **2**005)

$$GRS_{tot} = \sum (\frac{B_i}{|B_{\max}|}) GRS_i$$

첫 번째 단계에서 단일 염기 다형성과 매개변수 형질에 대한 회귀계수와 구분하기 위해 두 번째 단계에서는 i번째 매개변수 형질에 대한 회귀계수를 B_i 로 나타내었다. B_{Max}는 i개의 매개변수 형질 중 가장 큰 값을 가진 매개변수 형질의 회귀계수이다.

4장 실제 자료를 이용한 관상동맥질환 발생에 대한 유전적 기여도 예측 모형

4.1 관상동맥 질환 발생에 대한 유전적 기여도 예측 모형

4.1.1 심혈관계질환 유전체센터 자료

분석에 사용한 자료는 연세대학교 심혈관계질환 유전체 연구센터에서 실시하는 검진에 동의한 대상자들의 임상적 자료(CGC자료)를 이용하였다. 자료는 고혈압, 당뇨 등의 질병이 없는 대조군과 연세의료원 심장혈관센터에서 관상동맥 조영술을 시행받은 환자 중 관상동맥 협착이 있는 관상동맥질환자인 환자군으로 이루어 졌다.

본 연구에 사용된 환경인자는 나이, 흡연유무, 음주유무, 운동여부이며 질환의 위험요인(risk factor)으로 알려진 총 콜레스테롤(Total Cholesterol, Tchol), 고밀도 지단백콜레스테롤(High-density lipoproteins, HDL), 중성지방(triglyceride, Tg), 체질량 지수(Body Mass Index, BMI), 수축기 혈압(Systolic Blood Pressure, SBP), 이완기 혈압(Diastolic Blood Pressure, DBP), 글루코즈(Glucose)를 매개변수 형질로 사용하였다. (Pasternak et al. 1996)

분석 대상 유전자는 심혈관계 질환 유전체센터에서 수집된 9개의 염색체에 위치한 23개의 단일 염기 다형성들에 대해 조사하였다.

4.1.2 분석 대상 유전자 선택

단일 염기 다형성들과 양적형질인 매개변수 형질에 대한 상대적인 기여도(Contribution)를 평가하기 위해 분석 대상 유전자를 선택한다. 심혈관계 질환의 위험요인으로 알려진 Tchol, HDL, Tg, BMI, SBP, DBP, Glucose를 매개변수 형질로 설정하였으며 단일 염기 다형성들의 유전형 간의 차이여부를 검정 한다.

매개변수 형질과 관련 있는 유전자를 선택하기 위해 이용 가능한 9개의 유전자에 위치한 **2**3개의 단일 염기 다형성들에 대해 모두 조사 하였다. [표 **1**]은 분석에 사용한 대조군 544명, 환자군 767명으로 총 **1**311명으로 불안전한 자료이며 CGC자료의 단일 염기 다형성 변수 이름과 유전자 위치에 대한 표이다.

표 1 . 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 converting enzyme (perptidyl-dipeptidase A) 1	17q23
ACE_2	C-93T(C2547T)	angiotensin I converting enzyme (perptidyl-dipeptidase A) 2	17q23
ACE_6	14094(I/D)	angiotensin I converting enzyme (perptidyl-dipeptidase A) 6	17q23
ACE_7	G14480C	angiotensin I converting enzyme (perptidyl-dipeptidase A) 7	17q23
ACE_8	T849S(A14519G)	angiotensin I converting enzyme (perptidyl-dipeptidase A) 8	17q23
ACE_10	A22982G	angiotensin I converting enzyme (perptidyl-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
APM1_1	T45G	Adiponectin	3q27
APM1_2	G276T	Adiponectin	3q27
APOA5_2	T-1331C	apolipoprotein A-V	11q23
APOA5_5	C-1399T	apolipoprotein A-V	11q23
APOA5_6	G-1029A	apolipoprotein A-V	11q23
APOA5_7	G-12A	apolipoprotein A-V	11q23
CETP_2	C-629A	cholesteryl ester transfer protein, plasma	16q21
CETP_3	TAQ1B(G/A)	cholesteryl ester transfer protein, plasma	16q21
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
PON1	Q192R(A/G)	paraoxonase 1	7q21.3

위의 단일 염기 다형성들 중 **2**단계 모형 적용시 변수들의 독립성 가정을 만족시키기 위해 같은 염색체에 위치한 하나의 단일 염기 다형성을 선택한다. 분석 대상 유전자들을 선택하기 위해 모형에 같은 염색체에 위치한 단일 염기 다형성을 하나씩 추가시켜 Partial-F 값을 가장 크게 만드는 변수를 선택하고 Partial-F 검정을 하여 유의하면 모형에 포함한다. 반대로 유의하지 않으면 설명변수를 제거한다. 위의 과정을 반복하여 적절한 단일 염기 다형성을 모형에 포함하는 방법인 단계적 회귀분석(Stepwise regression Analysis)를 수행하였다. 단계적 회귀분석 결과는 아래의 [표 **2**]와 같다.

표 2 . 단계적 회귀분석(Stepwise regression) 결과

		Tchol			HDL			Tg			BMI			SBP			DBP			GLU		
		Beta	Partial R-Squ.	Sig.	Beta	Partial R-Squ.	Sig.	Beta	Partial R-Squ.	Sig.	Beta	Partial R-Squ.	Sig.	Beta	Partial R-Squ.	Sig.	Beta	Partial R-Squ.	Sig.	Beta	Partial R-Squ.	Sig.
ABCA1	GA				-3.37	0.01	0.03	18.54	0.01	0.08	0.45	0.01	0.08				-1.43	0.01	0.14	7.58	0.02	0.00
	GG				-2.27	0.01	0.12															
ACE_1	TT																					
	AT										-0.56	0.01	0.02									
ACE_2	TT	-13.41	0.01	0.05																		
	CT																					
ACE_6	ID							-259.93	0.02	<.0001										-35.34	0.02	0.01
	II	16.88	0.00	0.02																		
ACE_7	CG							209.42	0.03	0.00										45.02	0.01	0.00
	GG				-2.17	0.01	0.08															
ACE_8	AG							63.48	0.01	0.08												
	GG																					
ACE_10	GA																			-13.27	0.01	0.01
	GG																					
AGT_2	GA													2.49	0.00	0.12				-15.16	0.01	0.04
	GG																			-10.63	0.00	0.14
AGT_3	CA													4.20	0.01	0.01						
	CC																					
AGT_4	GA							16.47	0.01	0.11				4.26	0.01	0.01	2.26	0.01	0.02			
	GG																					
AGT_5	CT	5.39	0.00	0.14							0.39	0.00	0.13									
	TT	17.38	0.01	0.07																		
APM1_1	GT	5.45	0.00	0.12																		
	GG																					
APM1_2	GT																			-5.80	0.01	0.02
	TT				2.48	0.01	0.07	21.28	0.01	0.11												
APOA5_2	GA							32.66	0.04	0.03												
	GG							85.70	0.04	0.00												
APOA5_5	CT	-13.10	0.01	0.08																		
	CC							35.06	0.02	0.07	0.83	0.01	0.09									
APOA5_6	GA										1.13	0.01	0.09									
	GG																4.25	0.01	0.12			
APOA5_7	GA				-4.19	0.05	0.00															
	GG																					
CETP_2	CA																1.48	0.00	0.15			
	CC							-43.80	0.01	0.00	-0.65	0.01	0.03									
CETP_3	GA	-19.60	0.02	0.00	-2.50	0.01	0.13	-19.03	0.01	0.11												
	GG	-24.56	0.01	0.00	-3.69	0.01	0.03															
CETP_5	GA																					
	GG	-18.67	0.01	0.00				-37.68	0.01	0.01				-2.99	0.01	0.14						
LDLR	TT							82.41	0.01	0.04	2.26	0.01	0.02	10.65	0.01	0.07	6.50	0.01	0.10	16.64	0.01	0.11
	CT																					
LIPC_4	GG	-5.64	0.01	0.12				-29.98	0.02	0.00				-8.37	0.01	0.00	-2.28	0.01	0.02			
	GA													-6.35	0.01	0.02						
PON1	GG																					
	GA																					

위의 단계적 회귀분석 수행 후 많은 단일 염기 다형성들 중 각각 공통의 매개변수 형질에 영향을 미치는 높은 정보력을 갖는 다수의 단일 염기 다형성들로 이루어진 집합을 구성하였다. 가장 유사한 집합으로 매개변수 형질은 HDL과 Glucose이고, 유전자는 ABCA**1**-R**21**9K(G/A), APM**1**_**2**-G**2**76T, ACE7-G**1**4480C의 단일 염기 다형성들을 선정하였다. 즉 분석을 하기 위해 선택된 유전자는 공통의 매개변수 형질이나 그들의 집합에 대해 유의한 관련성이 있는 것이 포함되어져야 하며 단계적 회귀분석을 통해 모형에 포함된 단일 염기 다형성들에 대해 그들의 유의성을 가정한다.

4.1.3 분석 대상 유전자

유사한 집합으로 선정한 불안전한 자료에 대해 ABCA**1**-R**21**9K(G/A), APM**1**_**2**-G**2**76T, ACE7-G**1**4480C의 단일 염기 다형성과 HDL, Glucose의 매개변수 형질의 결측값과 이상값을 제외한 후 완전한 자료로 만들어 분석에 이용하였고 성별의 효과를 남자로 통제한 CGC자료의 분포는 정상군 84명(37%)으로 평균나이는 49±**12**.6**2**세, 흡연자 66명(34.0**2**%), 음주력 대상자 69(40.35%), 운동하는 자 37명(3**1**.09%)이며, 환자군은 **141**명(63%)로 평균나이는 56±7.**2**9세, 흡연자는 **12**8명(65.98%), 음주 대상인 자는 **102**명(59.65%), 운동하는 자는 **82**명(68.9**1**%)의 분포를 보이는 총 **22**5명으로 [표 3]과 같다.

표 3. 환경적 요인과 Intermediate Phenotype 분포

		Normal		CAD		Total	
		n	%	n	%	n	Mean±Std
Age(Mean±Std)		84	49.15±12.62	141	56.45±7.29	225	53.72±10.23
Smoke	No	18	58.06	13	41.94	31	
	Yes	66	34.02	128	65.98	194	
Drink	No	15	27.78	39	72.22	54	
	Yes	69	40.35	102	59.65	171	
Exercise	No	47	44.34	59	55.66	106	
	Yes	37	31.09	82	68.91	119	
		n	Mean±Std	n	Mean±Std	n	Mean±Std
HDL_C		84	44.43±8.83	130	37.03±8.39	225	39.68±9.24
Glucose		84	90.20±23.86	130	101.51±30.33	225	97.29±28.56

4.2 첫 번째 단계 : SNP Analysis 결과

먼저 선별된 유전자의 집합에 따라 각각의 유전자형에 대한 가변수(dummy variable)를 정의한 후 매개변수 형질에 유의한 효과가 존재하는지에 대해 아래의 선형회귀모형(Linear regression)을 설정할 수 있다.

$$IP = \beta_0 + \beta_1 G_{AA} + \beta_2 G_{AG} \quad (1)$$

이때, G_{AA}는 유전형이 AA면 **1**이고, 아니면 0의 값을 갖는 가변수이고 β_1은 그에 따른 회귀계수(Regression coefficient)이다. 마찬가지로 G_{AG}는 유전형이 AG이면**1**, 아니면 0의 값을 갖는 가변수이고 β_2는 그때의 회귀계수이다. 위의 회귀모형 (**1**)에서 회귀계수의 추정치(estimates)가 통계학적으로 유의하다면 유전적 영향이 존재한다고 평가한다.

위의 과정을 통해 구성된 후보 유전자들과 매개변수 형질들과의 관련성 분석(Association analysis)을 평가한다. 단일 염기 다형성들을 독립변수로 매개변수 형질과의 회귀분석을 수행하고, 환경변수와 매개변수 형질과의 회귀분석을 수행하여 얻은 회귀계수를 이용하여 후보 유전자의 기여위험도에 대한 점수(GRS_i)를 계산한다.

유전자의 영향과 환경적 요인이 매개변수에 유의한 효과가 존재하는지에 대한 모형은 다음의 식과 같이 표현할 수 있다.

$$\begin{aligned}\text{Intermediate Phenoytpe} \;=\; & \beta_0 + \beta_1 Age + \beta_2 Smoke + \beta_3 Drink + \beta_4 Exercise \\ & + \beta_5 ABCA1_{GA} + \beta_6 ABCA1_{GG} \\ & + \beta_7 ACE7_{CG} + \beta_8 ACE7_{GG} \\ & + \beta_9 APM12_{GT} + \beta_{10} APM12_{TT}\end{aligned}$$

2단계 모형의 첫 번째 단계인 SNP-Analysis에서는 위에서 보인 각각의 회귀분석 결과 얻은 비표준화회귀계수를 변화시킨 표준화회귀계수만큼 SNP_j에 가중치를 주어 후보 유전자의 영향력에 대한 기여위험도에 대한 점수를 구하는 단계이다.

4.3 두 번째 단계 : IP Analysis 결과

2단계 모형의 두 번째 단계인 IP(Intermediate Phenotype)-Analysis에서는 전체적인 경로의 유전자의 기여위험도(overall pathway's GRS, 이하 GRS_{tot})을 산출하고자 한다. GRS_{tot}은 독립변수인 매개변수 형질과 종속변수로 질환(D=0, D=**1**)에 대한 연관성을 평가하기 위해 로지스틱 회귀분석(Logistic regression analysis)을 수행한다. 그에 대한 회귀 모형식은 다음과 같이 나타낸다.

$$\log[P(D=1)] = B_0 + B_1 IP_{HDL} + B_2 IP_{Glu}$$

로지스틱 회귀분석을 통해 얻은 매개변수 형질에 대한 회귀계수 B_1 , B_2 중 가장 큰 값을 갖는 회귀계수를 $|B_{Max}|$으로 할때, 각각의 매개변수의 회귀계수를 $|B_{Max}|$값으로 나누어 준다. (표) 첫 단계에서 구한 GRS_i에 $\frac{B_i}{|B_{Max}|}$만큼의 가중치를 주어 전체 경로에 대한 유전자의 영향력을 구한 점수인 GRS_{tot}을 구해준다.

[표 4]는 HDL과 Glucose인 매개변수 형질과 단일 염기 다형성의 연관성을 평가하기 위해 선형 회귀분석 모형 결과이며 매개변수 형질에 대한 j번째 단일 염기 다형성의 가중치를 주는 회귀 계수를 정리한 표이다.

[표 4] 각각의 SNP와 매개변수의 선형회귀분석 결과

		ABCA1			ACE7			APM1_2		
IP	B	AA*	GA	GG	CC*	CG	GG	TT*	GT	GG
HDL	-0.099	0	-0.084	-0.**11**3	0	0.035	-0.0**2**3	0	-0.004	0.00**1**
Glucose	0.0**2**5	0	-0.008	0.**13**1	0	-0.084	-0.049	0	-0.0**2**8	-0.**1**09

(*는 기준범주)

4.4 Genetic Risk Score 분석

4.4.1 Genetic Risk Score 분석 결과

2단계 모형을 적용해 계산되어진 Genetic Risk Score은 관상동맥 질환을 예측하는데 있어 다인유전자 효과(Polygenic effect)를 조사하기 위한 독립변수로 사용할 수 있기 때문에 전체 경로의 Genetic Risk Score이 관상동맥 질환발생에 미치는 영향에 대해 로지스틱 회귀분석을 실행한다. 로지스틱 회귀모형은 다음과 같다.

$$\log\left(\frac{P(D=1)}{1-P(D=1)}\right) = \beta_0 + \beta_1 GRS_{tot} = \log\ odds$$

로지스틱 회귀 모형식을 이용하여 관상동맥 질환발생 확률을 구할 수 있다. 위의 식에 log를 제거하기 위해 지수(Exponential, 이하 Exp)를 취해주면 log를 제거할 수 있다. 따라서 원래의 양변에 지수를 Exp를 취해주면 왼쪽 항은 log가 제거되고 오른쪽 항은 다음과 같이 전환된다.

$$\frac{P(D=1)}{1-P(D=1)} = \mathrm{Exp}(\log\ odds) = odds$$

위의 식에 양변에 $1-P(D=1)$을 곱하여 주면 관상동맥 질환발생 확률인 $P(D=1)$을 구할 수 있다. 식은 다음과 같다.

$$P(D=1) = \frac{odds}{1+odds}$$

실제 자료의 구조상 샘플 수가 작아 위험인자가 낮은 대상자들로 이루어진 집단에서 단일 염기 다형성들이 나타낼 수 있는 가능한 모든 유전자형의 조합을 구성할 수 없었다. 실제 자료를 통해 얻어진 회귀 모형식을 적용하여 위험인자가 낮은 대상자들의 자료를 구성하여 모든 가능한 유전자형의 조합을 갖는 자료에 대입하여 단일 염기 다형성들의 효과를 고려한 질병발생 위험 확률을 구한다.

실제 자료로 구현된 로지스틱 회귀 모형식은 매개변수 형질이 HDL일 때, 관상동맥 질환 발생 확률의 모형식은

$$\log\left(\frac{P(D=1)}{1-P(D=1)}\right) = -0.028 + 6.4\,GRS_{tot}$$

$$P(D=1) = \frac{\mathrm{Exp}(-0.028 + 6.4\,GRS_{tot})}{1+\mathrm{Exp}(-0.028 + 6.4\,GRS_{tot})}$$

으로 나타내어 지고, 매개변수 형질이 Glucose일 때는

$$\log\left(\frac{P(D=1)}{1-P(D=1)}\right) = -0.005 + 7.652\,GRS_{tot}$$

$$P(D=1) = \frac{\mathrm{Exp}(-0.005 + 7.652\,GRS_{tot})}{1+\mathrm{Exp}(-0.005 + 7.652\,GRS_{tot})}$$

의 모형식으로부터 관상동맥 질환 발생 확률을 구할 수 있다.

실제 자료로 구해진 위의 로지스틱 회귀 모형식을 이용하여 환경적 인자의 상태가 질병발생에 영향을 주지 않을 경우 질병발생 위험에 기여하는 유전적 영향에 대한 전체 경로에 대한 Genetic Risk Score값의 변화에 따라 질병에 걸릴 확률을 계산해 낼 수 있다. [표 5]는 HDL 매개변수 형질에 의한 전체 경로에 대한 Genetic Risk Score값에 대한 3개의 단일 염기 다형성들의 독립변수들은 3개의 유

전형을 갖을 때 그들이 나타낼 수 있는 가능한 조합으로 총 **2**7개의 조합에 따른 관상동맥 질환발생 확률을 나타낸 표이다.

[표 5] HDL 매개변수에 대한 유전자형 조합에 따른 관상동맥 질환 발생 확률

Age>=45	Smoke	Drink	Exercise	ABCA1	ACE_7	APM1_2	Prob.(%)
No	No	No	Yes	AA	CG	GT	35.7
No	No	No	Yes	AA	CG	TT	35.8
No	No	No	Yes	AA	CG	GG	36.2
No	No	No	Yes	AA	CC	GT	43.9
No	No	No	Yes	AA	CC	TT	44.0
No	No	No	Yes	AA	CC	GG	44.4
No	No	No	Yes	AA	GG	GT	45.8
No	No	No	Yes	AA	GG	TT	45.9
No	No	No	Yes	AA	GG	GG	46.3
No	No	No	Yes	GA	CG	GT	48.8
No	No	No	Yes	GA	CG	TT	49.0
No	No	No	Yes	GA	CG	GG	49.3
No	No	No	Yes	GG	CG	GT	53.4
No	No	No	Yes	GG	CG	TT	53.5
No	No	No	Yes	GG	CG	GG	53.9
No	No	No	Yes	GA	CC	GT	57.4
No	No	No	Yes	GA	CC	TT	57.5
No	No	No	Yes	GA	CC	GG	57.8
No	No	No	Yes	GA	GG	GT	59.2
No	No	No	Yes	GA	GG	TT	59.3
No	No	No	Yes	GA	GG	GG	59.7
No	No	No	Yes	GG	CC	GT	61.8
No	No	No	Yes	GG	CC	TT	61.9
No	No	No	Yes	GG	CC	GG	62.2
No	No	No	Yes	GG	GG	GT	63.5
No	No	No	Yes	GG	GG	TT	63.6
No	No	No	Yes	GG	GG	GG	64.0

나이가 45세 이상이며, 흡연력과 음주력이 없으며, 운동을 하는 대상들에서 유전형의 조합에 따라 관상동맥 질환 발생확률의 차이가 있다. 즉 **2**단계 모형을 이용하여 질병발생 위험에 기여하는 유전적 영향에 대한 관상동맥 질환 발생 확률은 유전형이 조합에 따라 차이가 있음을 보인다.

[표 6]은 Glucose 매개변수 형질에 따른 전체 경로에 대한 Genetic Risk Score값에 대한 관상동맥 발생확률을 나타낸 표이다.

[표 6] Glucose 매개변수에 대한 유전자형 조합에 따른 관상동맥 질환 발생 확률

Age>=45	Smoke	Drink	Exercise	ABCA1	ACE_7	APM1_2	Prob.(%)
No	No	No	Yes	AA	CC	GG	35.4
No	No	No	Yes	AA	CC	TT	23.5
No	No	No	Yes	AA	CC	GT	35.8
No	No	No	Yes	AA	GG	GG	33.3
No	No	No	Yes	AA	GG	TT	21.9
No	No	No	Yes	AA	GG	GT	33.6
No	No	No	Yes	AA	CG	GG	31.5
No	No	No	Yes	AA	CG	TT	20.5
No	No	No	Yes	AA	CG	GT	31.9
No	No	No	Yes	GG	CC	GG	44.5
No	No	No	Yes	GG	CC	TT	31.1
No	No	No	Yes	GG	CC	GT	44.9
No	No	No	Yes	GG	GG	GG	42.2
No	No	No	Yes	GG	GG	TT	29.0
No	No	No	Yes	GG	GG	GT	42.5
No	No	No	Yes	GG	CG	GG	40.3
No	No	No	Yes	GG	CG	TT	27.4
No	No	No	Yes	GG	CG	GT	40.6
No	No	No	Yes	GA	CC	GG	35.2
No	No	No	Yes	GA	CC	TT	23.3
No	No	No	Yes	GA	CC	GT	35.5
No	No	No	Yes	GA	GG	GG	33.0
No	No	No	Yes	GA	GG	TT	21.7
No	No	No	Yes	GA	GG	GT	33.3
No	No	No	Yes	GA	CG	GG	31.3
No	No	No	Yes	GA	CG	TT	20.3
No	No	No	Yes	GA	CG	GT	31.6

매개변수 형질이 HDL을 통해 계산된 유전적 기여 위험도가 Glucose 매개변수 형질에 대한 유전적 기여 위험도에서 보다 관상동맥 질환 발생 확률을 대체로 더 높게 예측하고 있음을 볼 수 있다.

4.4.2 One-Stage 로지스틱 회귀분석

환경인자와 위험인자를 보정하여 단일 염기 다형성들과 질병과의 연관성 분석을 수행하는 기존의 방법과 **2**단계 모형을 적용한 결과를 비교한다.

[표 7]은 환경인자와 위험인자를 보정한 로지스틱 회귀분석 결과로 관상동맥 질환 발생확률을 나타낸 표이다. One-Stage 로지스틱 회귀분석결과 관상동맥 질환 발생에 미치는 유전적 기여위험은 거의 나타나지 않았다.

[표 7] 로지스틱 회귀분석결과 (Ajusted for Age, Smoke, Drink, Exercise, HDL)

Age>=45	Smoke	Drink	Exercise	ABCA1	ACE_7	APM1_2	Prob.(%)
No	No	No	Yes	GG	GG	GG	0.000000015
No	No	No	Yes	GG	GG	TT	0.000000005
No	No	No	Yes	GG	GG	GT	0.000000015
No	No	No	Yes	GG	CC	GG	0.000000019
No	No	No	Yes	GG	CC	TT	0.000000006
No	No	No	Yes	GG	CC	GT	0.000000020
No	No	No	Yes	GA	GG	GG	0.000000008
No	No	No	Yes	GA	GG	TT	0.000000003
No	No	No	Yes	GA	GG	GT	0.000000008
No	No	No	Yes	GA	CC	GG	0.000000010
No	No	No	Yes	GA	CC	TT	0.000000003
No	No	No	Yes	GA	CC	GT	0.000000010
No	No	No	Yes	GG	CG	GG	0.000000022
No	No	No	Yes	GG	CG	TT	0.000000007
No	No	No	Yes	GG	CG	GT	0.000000022
No	No	No	Yes	GA	CG	GG	0.000000012
No	No	No	Yes	GA	CG	TT	0.000000004
No	No	No	Yes	GA	CG	GT	0.000000012
No	No	No	Yes	AA	GG	GG	0.000000008
No	No	No	Yes	AA	GG	TT	0.000000003
No	No	No	Yes	AA	GG	GT	0.000000009
No	No	No	Yes	AA	CC	GG	0.000000011
No	No	No	Yes	AA	CC	TT	0.000000004
No	No	No	Yes	AA	CC	GT	0.000000011
No	No	No	Yes	AA	CG	GG	0.000000012
No	No	No	Yes	AA	CG	TT	0.000000004
No	No	No	Yes	AA	CG	GT	0.000000012

[표 8] 로지스틱 회귀분석 결과. (Ajusted for Age, Smoke, Drink, Exercise, Glucose)

Age>=45	Smoke	Drink	Exercise	ABCA1	ACE_7	APM1_2	Prob.(%)
No	No	No	Yes	GA	CC	TT	0.0000000246
No	No	No	Yes	GA	CC	GT	0.0000000247
No	No	No	Yes	AA	CC	TT	0.0000000278
No	No	No	Yes	AA	CC	GT	0.0000000279
No	No	No	Yes	GA	CC	GG	0.0000000299
No	No	No	Yes	AA	CC	GG	0.0000000337
No	No	No	Yes	GA	GG	TT	0.0000000339
No	No	No	Yes	GA	GG	GT	0.0000000341
No	No	No	Yes	GA	CG	TT	0.0000000367
No	No	No	Yes	GA	CG	GT	0.0000000368
No	No	No	Yes	AA	GG	TT	0.0000000383
No	No	No	Yes	AA	GG	GT	0.0000000385
No	No	No	Yes	GA	GG	GG	0.0000000412
No	No	No	Yes	AA	CG	TT	0.0000000414
No	No	No	Yes	AA	CG	GT	0.0000000416
No	No	No	Yes	GA	CG	GG	0.0000000445
No	No	No	Yes	AA	GG	GG	0.0000000465
No	No	No	Yes	GG	CC	TT	0.0000000492
No	No	No	Yes	GG	CC	GT	0.0000000494
No	No	No	Yes	AA	CG	GG	0.0000000502
No	No	No	Yes	GG	CC	GG	0.0000000597
No	No	No	Yes	GG	GG	TT	0.0000000678
No	No	No	Yes	GG	GG	GT	0.0000000681
No	No	No	Yes	GG	CG	TT	0.0000000733
No	No	No	Yes	GG	CG	GT	0.0000000736
No	No	No	Yes	GG	GG	GG	0.0000000822
No	No	No	Yes	GG	CG	GG	0.0000000889

One-Stage Model에 적용한 분석 결과 단일 염기 다형성의 효과만을 고려했을 때, 각각의 HDL, Glucose 의 매개변수 형질을 보정하여 분석한 결과에서 모두 질병 발생에 미치는 유전적 기여위험도를 찾기란 매우 어렵다는 것을 보였다. 따라서 위험인자를 갖고 있지 않는 대상에 대해 단일염기 다형성의 효과만을 봤을 때 **2**단계 모형을 적용한 결과에서 약 40배정도 관상동맥 질환 발생 예측 확률을 높일 수 있다.

5장 결론 및 고찰

유전자와 질병의 발생 사이에는 완벽한 연관성이 부족하기 때문에 환경적, 유전적 요인을 함께 고려해야 한다. 이러한 다요인성 질환은 아마도 해로운 요인의 수에 대한 어떤 개념적인 경계를 넘어 개인적인 결과로 인한 것일 수 있다. (Scheuuner M.T., 2004) 따라서 질병에 영향을 미치는 대부분의 위험 요인을 매개변수 형질로 정의 하였다. 즉 이러한 위험인자에 많이 노출된 사람은 질병의 위험도를 증가 시킬 것이며 하나의 매개변수 형질에 영향을 미치는 단일 유전자의 연관성에 대한 기존의 연구와 달리 공통의 매개변수 형질에 영향을 미치는 다수의 유전자들을 이용하면 그들의 정보력을 높일 수 있다. 이러한 점을 반영하여 지금까지 낮은 위험 요인을 갖는 대상에서 원인이 되는 유전적 효과를 찾기 위한 단일 염기 다형성들의 조합에 따라 관상동맥 질환에 미치는 유전적 기여위험도를 알아보기 위해 2단계 모형에 적용하여 평가하였다. 2단계 모형은 매개변수 형질을 고려하여 질병에 의미 있는 적용을 하기 위한 유전적 기여위험도(Genetic risk score)를 기술하는 방법이다. 이러한 전체 경로에 대한 유전적 기여 위험도는 매개변수 형질을 토대로 관상동맥 질환에 대해 유의한 위험 층화(Risk stratification)를 제공하며, 이 모형에 매개변수 형질을 고려하므로 유전자와 관상동맥 질환과의 원인적인 관계를 지지해 준다.

본 연구에서는 연세대학교 심혈관계 질환 유전체 연구센터에서 실시하는 검진에 동의한 대상자로 남자만을 통제한 자료를 가지고 관상동맥 질환 발생 예측 모형에 적용하였다. 나이가 45세 미만으로, 흡연력과 음주력이 없으며 운동을 하는, 대체로 위험인자가 낮은 경우에서 ABCA1-R219K(G/A), APM1_2-G276T, ACE_7-G14480C의 유전자들의 표현형에 대한 조합에 따라 질병 발생 위험에 기여하는 유전적 영향을 찾아내고자 하였다. 2단계 모형 적용으로 유전적 효과 자체만을 고려했을 때 약 30%이상의 질병에 미치는 유전적 영향력을 보일 수 있었다. 반면에 기존의 환경인자와 위험인자들의 보정하여 유전적 영향의 연관성을 찾고자 했던

One-Stage Model에서는 유전적 영향력의 정도를 찾아내기란 쉽지 않았다. 즉 기존 분석 방법과 달리 질병 발생 위험에 기여하는 유전적 영향을 찾기 위한 **2**단계 모형의 적용은 보다 더 잘 접근할 수 있는 방법임을 보였다. 또한 관상동맥 질환에 HDL 매개변수 형질이 Glucose 매개변수 형질일 때 보다 유전적 효과에 대한 기여도를 더 잘 예측하였다. 실제로 수많은 방법과 과정으로 얻어진 표본은 연관성 분석 과정에서 매우 축소되기 마련이다. 이러한 자료의 구조상 관상동맥 질환 발생 확률 예측을 과추정한 경향이 있지만 유전적인 기여위험을 더 잘 밝힐 수 있다는 점을 보일 수 있다.

2단계 모형을 통해 관상동맥 질환은 매개변수 형질과, 유전자형-매개변수 형질의 관계에 영향을 받을 것이며 매개변수 형질과 질환의 변이성은 다형성의 유사한 집합에 의해 부분적으로 영향을 받을 것이다. 또한, 모형에 적용하기 위해서는 분석 대상 단일 염기 다형성의 수가 많아야 하며, 표본수가 많아야 한다는 제한점을 갖고 있다. 매개변수 형질의 수가 많아지면 해석에 어려움이 생기며 공통의 매개변수 형질에 영향을 미치는 다수의 단일 염기 다형성의 교집합을 구성하기가 힘들다. 이러한 점은 연구계획과 밀접한 관계가 있다. 자료 수집시 단일 염기 다형성의 유전자형과 매개변수 형질의 관측값과 환경적 요인에 대한 모든 정보를 얻어야 한다.

본 연구에서 **2**단계 모형의 적용으로 관상동맥 질환에 대한 Genetic Risk Score는 중요한 임상적 영향력을 보이는 지표로서 관상 동맥 질환의 다양한 위험요인을 반영하며 구체적인 평가 방법임을 확인하였다. 이러한 모형 적용을 통해 보여진 단일 염기 다형성들의 조합이 질병발생에 미치는 영향을 구체적으로 규명하는 것은 매우 의미 있는 일이라 판단된다.

참 고 문 헌

Akcay, A., Sezer, S., Ozdemir, F.N., Arat, Z., Atac, F.B., Verdi, H., Colak, T., Haberal, M. (2004). Association of the Genetic Polymorphisms of the Renin-Angiotensin System and Endothelial Nitric oxide Synthase With Chronic Renal Transplant Dysfunction. Transplantation 78: 892-898

Carlson, C.S., Eberle, M.A., Kruglyak, L. and Nickerson, D.A. (2004). Mapping complex disease loci in whole-genome association studies. Nature 429:446-452

Cheng, S., Grow, M.A., Pallaud, C., Klitz, W., Erlich, H.A., Visvilis, S., Visvikis, S., Chen, J.J., Pullinger, C.R., Malloy, M.J., Siest, G., and Kane, J.P. (1999). A Multilocus Genotyping Assay for Candidate markers of Cardiovascular Disease Risk. Genome Res 9: 936-949

Donna K., Arnett, J.S., Pankow, L.D., Atwood, and Thomas A. Sellers. (1997). Impact of Adjustments for Intermediate Phenotypes on the Power to Detect Linkage. Genet Epidemiol 1997; 14:749-754

Fiona R. Green : Genetic risk markers for post-angioplasty restenosis: What should we expect?. School of Biomedical and Molecular Sciences, University of Surrey, Guildford GU2 7XH, UK.

Genest, J., and Cohn, J.S. (1995). Clustering of cardiovascular risk factors: targeting high-risk individuals. Am J Cardiol 76:8A-20A

Geral, F. and Dibona. (1999). Renal Mechanoreceptor Dysfunction An Intermediate penotype In Spontaneously Hypertensive Rats. Hypertension. 33:472-475

Hosmer, D.W., and Lemeshow, S. (2001). Applied Logistic Regression 2nd ed. New york: John Wiley & Sons

Horne, B.D., Anderson, J.L., Carlquist, J.F., Muhlestein, J.B., Renlund, D.G., Bair, T.L., Pearson, R.R., and Camp, N.J. (2005). Generating Genetic Risk Scores from Intermediate Phenotypes for Use in Association Studies of Clinically Significant Endpoints. An of Hum Genet 69:176-186

Hoeg, J.M. (1997). Evaluating coronary heart disease risk. J. Am. Med. Assoc. 277:1387-1390

Hischhorn, J.N., Lohmueller, K., Byrne, E., and Hirschhorn, K. (2002). A comprehensive review of genetic association studies. Genet Med 4(2):45-61

Hamon, S.C., Kardia, S.L., Boerwinkle, E., Liu, K., Klos, K.L., Clark, A.G., and Sing, C.F. (2006). Evidence for Consistent Intragenic and Intergenic Interactions between SNP Effects in the APOA1/C3/A4/A5 Gene Cluster. Hum Hered 61:87-96

Lange, K. Mathematical and Statistical methods for Genetic Analysis. 2nd ed. Springer; 2002.

Nitsch, D., Molokhia, M., Smeeth L., DeStavola, B.L., Whittaker, J.C., and Leon, D.A. (2006). Limits to Causal Inference based on Mendelian Randomization: A Comparison with Randomized Controlled Trials. Am J Epidemiology. 1;163(5):397-403

Pharoah, P. D. P., Antoniou, A., Bobrow, M., Zimmern, R. L., Easton, D. F. and Ponder, B. A. J. (2002). Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet 31:33-36

Sander, Greenland. (2000). An introduction to instrumental variables for epidemiologists. International journal of Epidemiology 29:722-729

Scheuuner, M.T. (2004). Clinical application of genetic risk assessment strategies for coronary artery disease: genotypes, phenotypes, and family history. Prim Care. 31(3):711-37

Vittinghoff, E. Regression, Methods in Biostatistics: linear, logistic, survival, and repeated measures models. New York: Springer; 2005.

Pasternak, R.C., S.M. Grundy, D. levy, and P.D. Thompson. (1996). 27th Bethesda Conference: Matching the intensity of risk factor management with the hazard for coronary disease events. Task Force 3. Spectrum of risk factors for coronary heart disease. J. Am. Coll. Cardiol. 27:978-990

Risch, N.J. Searching for genetic determinants in the new millennium. Nature VOL 405 15 JUNE 2000 (www.nature.com)

Rozek, L.S., Hatsukami, T.S., Richter, R.J., Ranchalis, J., Nakayama, K., McKinstry, L.A., Gortner, D.A., Boyko, E., Schellenberg, G.D., Furlong, C.E., Jarvik, G. (2005). The correlation of paraxonase (PON1) activity with lipid and lipoprotein levels differs with vascular disease status. J Lipid Research 46

Winkelmann, B.R., Hager, J., Kraus, W.E., Merlini, P., Keavney, B., Grant, P.J., Muhlestein, J.B., and Granger, C.B. (2000). Genetics of coronary heart disease: current knowledge and research principles. Am Heart J 140:S11-S26

ABSTRACT

Prediction for the Genetic Contribution in Coronary Artery Disease Using Two-Stage Model

Lee, Hye Jin.

Dept. of Biostatistics and Computing

The Graduate School

Yonsei University

Coronary artery disease(CAD) was known as a complex disease resulting from mutation of multiple related genes and established risk factors such as age, BMI, smoking, drinking and lipid profiles and etc. There exist some approaches such as one-stage model that could identify genetic risk factor for a disease only through direct analysis of candidate genes, usually through association studies. However, genetic susceptibility to CAD was not addressed adequately by these methods.

In this thesis, we proposed using the two-stage model approach to calculate a genetic risk score for a coronary artery disease from many candidate genes and multiple intermediate phenotypes. In first stage, linear regression analysis was used to evaluate the association of candidate genes and environmental

factors with intermediate penotypes. And then, logistic regression analysis was used to assess the overall genetic contribution to the risk of coronary artery disease in second stage.

The results of analysis of two-stage model approach suggested that the genetic risk score could be widely used for assessment of genetic risk to contribute to the development of CAD.

Key words : Two-Stage Model, Intermediate Phenotypes, Linear regression, Logistic Regression, Genetic contribution.