유전적 연관성 분석을 위한 가계도 자료의 정보 충분성에 관한 연구

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
박 윤 주

유전적 연관성 분석을 위한 가계도 자료의 정보 충분성에 관한 연구

지도 김 동 기 교수

이 논문을 석사 학위논문으로 제출함

2001년 6월 일

연세대학교 대학원
의학전산통계학협동과정
의학통계학전공
박 윤 주

박윤주의 석사학위 논문을 인준함

심사위원 _____인

심사위원 _____인

심사위원 _____인

연세대학교 대학원

2001년 6월

감사의 글

어느덧 삼복 무더위가 기승을 부리는 여름의 한가운데 이르렀습니다. 대학원 생활 처음 시작할 때의 두려움과 설레임을 뒤로하고 벌써 졸업의 시기에 이르러서야 새삼 세월의 빠름을 실감합니다.

그 동안 저의 부족함에도 불구하고 끊임없는 충고와 질책으로 대학원 생활은 이끌어 주신 김동기 교수님께 감사드립니다. 바쁘신 시간 중에도 논문을 위해 아낌없는 도움을 주신 임길섭 교수님께도 감사의 말을 올립니다. 대학생활에서 통계학에 대해 알게 해 주신 김동건 교수님 조진남 교수님께도 감사를 드립니다.

대학원 생활 처음 시작할 때 서투른 저를 이끌어 주신 지연언지, 희중오빠, 기준오빠, 종건선배, 항상 편안한 웃음으로 반겨주시는 강대룡 선생님, 항상 밝은 모습으로 모두를 즐겁게 해 주었던 경민이와 효미 연구실에 있으면서 많은 도움을 받았던 성민 오빠, 덜렁거리지만 속 깊은 은정, 큰언니처럼 푸근한 현선언니, 명희언니, 찬미, 영선, 무영, 현철, 성재씨에게 감사의 말을 전합니다. 대학원 생활동안 함께 동고동락한 나의 동기 시내언니, 희철오빠, 우선오빠, 영진이 에게도 고맙다는 말을 전합니다.

힘들었지만 2년간의 생활을 통해 많은 것을 얻은 지금 그 섭섭함과 아쉬움을 가지고 사회 속으로 한 발 내딛으려 합니다. 여러 선후배와의 만남과 짧은 시간이었지만 김미정 박사님과의 만남을 인생에서의 소중한 추억으로 간직하고 싶습니다.

언제나 어려울 때 옆에 있어 주었던 나의 가장 친한 벗 영림이, 은경이 친척 동생 소현이 그 고마움을 말로는 다 표현하지 못하지만 이 지면을 빌어 마음을 대신합니다.

항상 친구처럼 때론 언니처럼 따스함을 주었던 나의 친구들에게도 고맙다

는 말을 하고 싶습니다.

긴 시간동안 한결같은 사랑의 마음으로 키워 주셨던 사랑하는 부모님께 감히 감사하다는 말을 전합니다. 힘드신 상황 속에서도 항상 옆에서 위로해 주셨던 아빠, 곁에는 없지만 지금 나의 모습을 가장 자랑스럽게 생각하실 엄마, 그리고 속 깊은 내 사랑하는 동생 현규, 항상 부족한 딸이고 누나였지만 언제나 따뜻하게 감싸주었던 나의 가족들이 있었기에 제가 지금 이 자리에 서 있을 수 있었습니다.

마지막으로 항상 내가 힘들 때마다 나를 붙들어 주시고 위로해 주셨던 그 분께 가장 큰 감사를 올립니다.

2001년 여름에 윤주 올림

제 목 차 례

표 차 례

그 림 차 례

국문요약

유전적 연관성 분석을 위한 가계도 자료의 정보 충분성에 관한 연구

통계적 검정력을 통한 가계도 자료의 정보 충분성 검정은 연관성분석에서 표식유전자 판독 전에 수집된 가계도의 정보의 충분성을 알아볼 수 있기 때문에 표식유전자 판독에 필요한 장기간의 시간 및 막대한 비용의 손실을 최소화 할 수 있다. 또한, 필요한 가계도의 수나 질병유전자와 관련있는 적절한 표식유전자를 예측할 수 있으므로 연구의 디자인 설계시 매우 유용하다.

본 논문에서는 이러한 정보 충분성 검정에 관한 이론을 정립하고 이를 위하여 DNA 및 유전자의 개념과 연관성 검정에서 발생하는 불완전 침투에서의 유전자 재조합률과 침투율의 추정방법을 정리하고 정보충분성 측정방법을 서술한다.

또한 정립된 이론을 실제자료에 적용하여 조사된 가계도의 정보 충분성을 추정하였다. 이를 위하여 일 병원에 내원한 심혈관질환(Cardiovascular disease; CVD)발생환자의 가계도를 사용하였다. 심혈관질환을 가지고 있는 2명의 환자를 대상으로 그들 가족들의 심혈관질환 발병여부를 조사하였는데, 첫 번째 가족은 27명중 심혈관질환을 가지고 있는 사람이 8명이었고 두 번째 가족은 17명 중 심혈관질환을 가지고 있는 사람은 4명이었다.

이 두 가계도를 대상으로 하여 정보의 충분성을 Expected maximum LOD score와 maximum LOD score가 특정 상수보다 클 확률을 사용하여 추정하였다. 그 결과 단일 표식유전자를 사용한 분석과 두 개의 표식유전자를 사용

한 추정 모두에서 첫 번째 가계도가 두번째 가계도 보다 검정력이 높았고, 단일 표식유전자를 이용한 분석에서는 질병유전자와 표식유전자의 거리가 근접할 수록, 두 개의 표식 유전자를 사용한 분석에서는 질병유전자가 두 표식 유전자 사이에 있다는 가정 하에서 두 표식유전자의 거리가 근접할 수록 가계도의 검정력이 높게 나왔다. 또한 동일한 가계도를 이용하였을 경우 단일 표식유전자를 이용한 분석보다는 두 개의 표식유전자를 이용한 분석에서의 가계도 자료에 대한 검정력이 높았고, 표식유전자의 두 대립형질의 빈도가 동일할수록 검정력이 높게 나오는 결과를 도출하였다.

핵심되는 말 : 정보충분성, 가계도 자료의 검정력, 연관성분석, 불완전침투, 재조합률, 침투율

제 1 장 서 론

1. 1 연구배경

2001년 2월 인간 게놈 프로젝트(HGP: human genome project)와 셀레라사(Celera corp.)의 연구결과가 발표되었는데, 이는 그동안 진행해온 인간 유전자 연구에 대한 중간발표로서, 약 32억개로 추산되는 인간의 DNA중 약 99%에 해당하는 부분을 판독하였음을 공표하고, DNA 서열(DNA sequence) 중 일부에 분포되어 있는 유전자(gene)는 26,000 40,000 정도로서 그동안 추산되어온 수보다는 작을 것임을 발표한 것이었다. 이러한 연구결과로 인간 DNA의 구성내용 식별은 거의 완성되었으며, DNA 서열에서의 유전자 위치 파악도 조만간 완료될 것임을 의미하는 것이다.

이와 함께 인간 DNA 서열을 데이터베이스 형태로 공개함으로써, 유전학 분야의 연구가 새로운 전기를 맞고 있다. 즉, 세계 각국에서는 인종별 DNA의 상이성 분석과 인간이외 동식물의 DNA 판독 등에 관한 연구를 가속화하고 있으며 이를 기반으로 한 생명체의 유전현상 규명과 질병치료법 개발 혹은 동식물 배양 등 다양한 분야에서의 연구가 폭발적으로 증가하고 있다.

그러나 금번 인간 DNA 서열 발표가 유전학 분야의 연구완료를 의미하는 것은 아니다. 현재까지의 성과는 인간 DNA 서열 자체와 상당수 유전자의 위치정도를 파악한 것으로서, 보다 중요한 사항인 유전자별 기능 식별에 관해서는 아직 초보단계라 할 수 있다. 즉 DNA 서열상의 유전자와 그것이 관여하는 인간의 특성(trait)을 연관(linkage)시킬 수 있는 경우에 유전현상의 규명 및 질병치료 등이 본격화 될 수 있으나, 아직은 대부분의 유전자 기능이 알려져 있지 않다. 따라서 인간 유전체(genome)에 관한 연구는 이제 시작단계에 있다고 할 수 있으며, 향후 연구초점중의 하나는 유전자의 위치와 그

기능의 연관성 파악이 될 것으로 예측되고 있다.

가계도(pedigree) 자료를 통한 연관성 분석은 이러한 유전자 기능 식별을 위한 연구의 일환이다. 이는 이미 위치가 알려져 있는 유전자(표식유전자: marker)를 활용하여 우리가 규명하고자 하는 특성에 관련된 유전자를 추론하는 것으로서, 가계도 구성인원의 특성과 표식유전자의 판독(유전자 타이핑: genotyping)을 통하여 수행된다. 이러한 분석에서는 다수의 가계도에 대한 많은 표식유전자의 판독이 요구되고, 이로 인하여 장기간의 시간 및 막대한 비용이 소요되는 것이 일반적이다. 따라서 가계도 자료를 통한 연관성 분석에서는 미지의 유전자에 관한 추론을 위하여 필요한 가계도 수준(수와 각각의 구성)과 표식유전자의 유형을 적절하게 설정하는 것이 중요한 문제로 대두된다. 이는 표식유전자의 적절성이 연관성 분석의 효율성에 영향을 미침은 물론이고, 과다한 가계도 조사는 시간 및 경제적인 측면에서 비효율적이며 적은 수의 가계도로는 원하는 수준의 추론을 수행할 수 없기 때문이다.

이러한 관점에서, 연관성 분석을 위하여 설정된 가계도 수준과 표식유전자의 적절성을 평가하는 중요한 척도로서 가계도 자료의 정보충분성(sufficiency of information)이 사용될 수 있다. 이때 정보충분성은 조사된 가계도가 함유하는 정보량을 지칭하며, 이의 측정은 가계도 구성원의 표식유전자 판독전에 수행된다.

따라서 정보충분성 측정결과를 토대로 가계도 구성원의 표식유전자 판독 필요성을 결정할 수 있으며, 이는 연관성 분석에의 시간 및 비용의 상당부분 혹은 대부분이 표식유전자 판독에 소요됨을 고려할 때 연관성 분석연구의 효율성을 좌우하는 중요한 요소라 할 것이다.

1. 2 연구목적 및 방법

본 연구는 연관성 분석시 가계도 자료가 조사된 경우 각 개인의 표식유전자 판독전에 조사된 가계도 자료와 선정된 표식유전자가 미지의 유전자 추론에 적절한지의 여부를 판단하는데 필요한 정보충분성을 측정하는 방법을 정립하고 이를 실제 한국에서 진행되고 있는 연관성 분석연구에 적용하여 분석하는데 그 목적을 둔다.

정보충분성 측정을 위하여 가계도 구성 및 표식유전자 유형이 전제되는 경우의 가계도가 함유하는 정보량의 추정과, 이때 연관성 검정을 실시할 경우의 검정력(power of test) 도출을 기본 관점으로 한다.

이를 위하여 정보충분성 측정을 위한 관련이론은 문헌조사를 통하여 정리하고, 실제적 연관성 분석연구에 활용되는 모의실험(simulation)기법을 SIMLINK 프로그램을 중심으로 조사한다. 또한 이러한 기법을 한국의 심혈관 질환의 유전적 연구를 위하여 수집된 가계도에 적용함으로서, 선정된 가계도와 표식유전자의 적절성을 분석한다. 이때 가계도 자료의 일반적 특징을 분석하기 위하여 SAS package를 사용한다.

1. 3 연구내용

본 논문은 정보충분성 측정을 위한 관련이론과 모의실험 기법의 정리 및 실제 자료에의 적용부분으로 구성되었다.

관련이론 부분은 정보량 추정과 검정역 도출에 필요한 DNA 및 유전자의 개념, 연관성 검정을 위한 검정통계량 설정방법, 연관성 검정에서 발생하는 불완전침투(incomplete penetrance)의 처리방법, 가계도의 정보량 추정과 검

정력 도출방법 등으로 구성되었다.

모의실험 기법의 정리 부분에서는 SIMLINK 프로그램을 중심으로 관련된 이론과 알고리듬을 정리하였고 실제 자료 적용부분에는 한국 심혈관 연구를 위하여 수집된 가계도 자료에 정보충분성 측정기법을 적용하여 분석한 결과를 제시하였다.

제 2 장 이론적 배경

2.1 개 요

정보충분성은 연관성 분석(linkage analysis)에서의 검정력을 기초로 추정하며, 연관성 분석에서는 우리가 알고자 하는 특성유전자의 위치(trait locus)를 이미 알려져 있는 표식유전자로부터의 거리로 추정하는 것을 주요 관점으로 한다. 이때의 표식유전자와 특성유전자간의 거리를 확률적 의미로 환산한 것을 재조합률(recombination fraction)이라 정의하고, 연관성 분석에서는 이 재조합률의 추정과 검정을 주요 목적으로 한다. 일반적으로 사용되는 연관성 분석시의 귀무가설은 두 유전자간 연관성이 없는 상황을, 대립가설은 연관성이 있는 상황으로 설정되며, 이러한 가설검정을 위하여 우도비(likelihood ratio)로 구성된 검정통계량이 사용된다.

따라서 정보충분성은 조사된 가계도에서 연관성이 있음을 전제할 때, 앞의 검정통계량을 사용하여 귀무가설을 기각할 확률로서 측정할 수 있으며 혹은 검정통계량의 확률적 기대값으로서도 측정될 수 있다.

제 2장에서는 정보충분성 측정의 기본단위가 되는 재조합률의 추정과 이를 위한 가설검정에 관련된 이론을 먼저 정리한 다음 정보충분성 측정방법을 서술한다. 특히 재조합률 추정시의 일반적 상황인 불완전 침투(incomplete penetrance) 경우의 관련이론도 정리한다.

그런데 정보충분성에 직접 관련된 여러 이론을 서술하기에 앞서 DNA 및 유전자의 개념부터 정리하고자 한다. 이는 연관성 분석에서 유전자 개념의 이해가 필수적으로 선행되어야 하기 때문이다.

2. 2 DNA 및 유전자의 개념

부모로부터 자식에게 특성이 전해지는 것은 유전자(gene)라고 부르는 물질에 의한 것인데, 유전자의 개념은 완두콩을 사용한 그레고리 멘델의 실험에 의해서 1900년대 초기에 제안되었다. 그러나 이때까지만 해도 유전물질은 단백질(protein)이라고 알려져 있었다(이정우, 1994).

1953년 왓슨과 크릭이 데옥시리보핵산(deoxyribonucleic acid : DNA)에 X선을 방사시켜 회절되어 나타나는 모양을 이용하여 생물체의 유전전달 물질인 DNA의 구조를 밝혔는데 이를 통해 유전전달 물질에 대한 연구는 새로운 전기를 맞이하게 되었다 (정해영 ,2000).

염색체(chromosome)위에 있는 DNA는 아데닌(A), 티민(T), 시토신(C), 구아닌(G)의 4가지 염기가 나열된 이중 나선 구조(double helix)로 되어 있다. 4가지 염기가 3개씩 조합된 코돈(codon)이 아미노산(amino acid)을 만들고 아미노산이 단백질(protein)을 형성한다. 유전정보의 전달은 DNA 단위로부터 1차 산물인 RNA로 전사(transcription)되며, 다음은 RNA로부터 단백질을 합성하기 위해 mRNA가 만들어지는데 이 과정을 잇기(splicing)라 한다. 또 mRNA가 2차 산물인 단백질을 합성하는데 이를 번역(translation)과정 이라 한다.

유전자는 이러한 DNA 서열(DNA sequence)중에서 일부분을 일컫는 단위이다. 즉, DNA는 유전적 기능을 하는 엑숀(exon)과 유전적 기능을 하지 않는 인트론(intron)부분으로 구성되어 있는데 우리가 일반적으로 유전자라 하는 부분은 이중 유전적 기능을 하는 엑숀 부분이다 (박상규 외, 1996 ; 이철훈, 1993; 강동민, 1995).

2. 3 연관성의 추정과 검정

2. 3. 1 연관성 분석의 개념

연관성 분석(linkage analysis)법은 2개 혹은 그 이상의 유전자가 서로 독립적으로 분리되는지 혹은 유전자들이 매우 근접하게 위치하기 때문에 독립적으로 분리되지 않고 연관되어 나타나는 지를 조사하여 유전자들간의 거리를 측정하는 방법이다. 세포 분열과정 중 염색체의 교차(cross over)로 인한 염색체 분절(segment)의 교환으로 새로운 조합의 대립유전자(allele)를 가지게 되는데 이를 유전자 재조합(recombination)이라 한다. 교차는 연관 유전자들을 재조합하여 부모에서는 발견되지 않은 새로운 대립유전자의 쌍을 만들어 낸다.

이 유전자 재조합은 두 유전자 사이의 거리가 가까우면 교차를 일으키는 빈도수가 낮아 재조합률(recombination fraction)이 낮아지고 거리가 멀면 유전자 사이의 교차 빈도가 높아져서 재조합률이 높아진다. 이와 같이 두 유전자의 연관(linked)정도와 유전적 거리(map distance)는 밀접한 관련이 있는데 이러한 연관 분석법을 이용하여 우리가 알고 싶어하는 특성 유전자의 위치(trait locus)를 찾아내는 것이 연관성 분석의 가장 큰 목적이다 (Jurg ott 1991).

2. 3. 2 재조합률의 추정과 검정

주어진 가계도 정보로부터 유전자 재조합률은 크게 우도 함수(likelihood function)와 베이지안(Bayesian)방법을 이용하여 추정(estimate)할 수 있다.

첫번째로 우도함수를 이용한 추정은 다음과 같다.

조사된 가계도의 정보를 F라 하고 두 유전자 사이의 재조합률을 θ 로 정

의하면 우도 함수는 L(θ | F)와 같이 표현 할 수 있는데, 이 함수는 가계도에 대한 정보가 주어졌을 때의 재조합률 θ의 확률에 의해 정의되어질 수 있다.

두번째로 베이지안 방법에 의한 재조합률 추정과정은 다음과 같다.

우선 θ 의 사전분포(prior distribution)를 미리 가정한 다음 주어진 가계도 자료의 표본 분포를 통해 L(F |θ) 를 계산하고 마지막으로 이 두 정보를 이용하여 다음과 같이 θ의 사후 분포(posterior distribution)를 통한 θ의 재조합률을 추정할 수 있다.

$$P(\theta \mid F) = \frac{L(F \mid \theta) h(\theta)}{\int L(F \mid \theta) h(\theta) d\theta}$$

$h(\theta)$: θ 의 사전분포

$P(\theta \mid F)$: θ 의 사후분포

위와 같이 두 유전자의 재조합률 추정 후 이에 대한 연관성 검정(linkage test)이 이루어지는데 일반적 연관성 검정의 가설은 아래와 같다.

H_0 : 두 유전자는 연관되어 있지 않다. ($\theta = \frac{1}{2}$)

H_1 : 두 유전자는 연관되어 있다. ($\theta < \frac{1}{2}$)

이 가설을 검정하기 위하여 일반적으로 Lod Score가 사용되는데 Lod Score는 아래와 같이 연관이 되어있지 않다고 가정했을 때의 우도 $L(\theta = \frac{1}{2} \mid F)$ 와 연관되어 있다고 가정한 우도 $L(\theta \mid F)$ 비율(proportion)의 log 값이다.

$$Z(\theta) = \log_{10} \frac{L(\theta \mid F)}{L(\theta = \frac{1}{2} \mid F)}$$

위 식에서 θ를 최대로 하는 값을 최대 Lod Score(maximum lod score)라고 하는데 아래와 같이 표현할 수 있다.

$$Z_{max} = \max . LOD = \log_{10} \frac{L(\hat{\theta} \mid F)}{L(\theta = \frac{1}{2} \mid F)}$$

$\hat{\theta}$: 최대우도 추정치 (maximum likelihood estimate : m.l.e.)

이러한 검정통계량은 표본 크기가 클때 귀무가설 하에서 자유도(degree of freedom)가 1인 χ^2 분포를 따르고 보통 $Z_{max} \geq 3$ 이면 연관되어 있지 않다는 귀무가설($H_0 : \theta = \frac{1}{2}$)을 기각 할 수 있다 (Lander and Bostein, 1989). 즉, 두 유전자가 연관되어 있다고 볼 수 있다.

2.4 불완전 침투에서의 연관성

2.4.1 불완전침투의 정의

일반적으로 연관성분석은 유전자의 유전형과 표현형이 동일하게 나타나는 것을 전제로 하고 있다. 즉, 인간의 경우 질병유전자(disease gene) D가 우성형질(dominant)일 때 병에 걸린(affected) 사람의 유전형은 Dd나 DD 가 되야 하고, 병에 걸리지 않은(unaffected)사람의 유전형은 반드시dd 가 되야 한다.

또한 D가 열성형질(recessive)인 경우는 DD인 경우가 병에 걸린 사람의 유전형이 되어야 할 것이다.

그러나 유전학적 관점으로 볼 때 개체가 질병을 일으키는 유전형을 가지고 있다 해도 발현되지 않는 경우 즉, 다시말해 질병유전자를 가지고 있다고 하더라도 병이 발병되지 않는 경우가 있는데, 이를 불완전 침투(incomplete penetrance)라 한다. 이와 반대로 질병유전자를 가지면 반드시 발병되는 경우를 완전 침투(full penetrance)라 한다 (Jurg ott 1991).

또, 침투율(penetrance 또는 penetrance rate)은 다음과 같이 유전형이 주어졌을 때 표현형의 조건부 확률로 정의 될 수 있는데 ($P(x \mid g)$, x: 표현형, g: 유전형) 완전침투는 질병유전자를 가지고 있으면 반드시 발병이 되므로 이 확률이 1이 된다. 그러나 불완전 침투의 경우 질병유전자를 가지고 있더라도 모두 발병이 되는 것은 아니기 때문에 침투율이 항상 1 보다 작아 진다.

그러므로 연관성 분석시 불완전 침투의 경우는 재조합률과 침투율을 모두 추정해야 한다.

2. 4. 2 우도함수를 사용한 재조합률과 침투율의 추정과 검정

일반적으로 침투율 λ와 재조합률 θ는 결합 우도(joint likelihood) $L(\lambda, \theta)$에 의해 추정한다. 이때 대부분 표식유전자의 정보를 이용하여 각 개인의 유전형을 알아내지만 부모의 유전형이 불확실하다면 침투율 λ와 재조합률 θ의 추정(estimation)은 (λ, θ) 의 여러가지 조합중 $L(\lambda, \theta)$ 을 최대화(maximize)시키는 조합으로 추정한다 (Rossen, R. D. et al, 1980).

침투율 λ는 보균자(질병 유전자를 보유한 사람, carrier)중에서 실제로 병에

걸린 사람의 비율로 추정되는데, 일반적으로 보균자를 쉽게 찾을 수 없기 때문에 불편의(unbiased) 추정을 하기 어렵다.

이 침투율의 추정에 관해서는 현재 다양한 연구가 진행되고 있다(Easton et al, 1995; Gail et al, 1999; Liang et al, 1999; Moore et al, 2000).

일반적으로 연관성 검정을 할 때 최대 LOD Score를 사용하는데 불완전 침투일 경우 최대 LOD Score는 다음과 같이 수정되어 나타낼 수 있다.

$$Z_{max} = Z(\hat{\theta}) = \log_{10} \frac{L(\lambda, \hat{\theta})}{L(\lambda, \theta = \frac{1}{2})} \qquad \hat{\theta} : \text{M.L.E}$$

위와 같은 Z_{max}는 침투율 λ를 알고 이 λ와 재조합률 θ가 독립(independent)일 때는 가능하다. 그러나 실제적으로 침투율은 대체적으로 추정되어지고 이 추정치는 재조합률과 상관관계(correlation)를 갖기 때문에 침투율 λ와 재조합률 θ는 (λ, θ)의 조합을 최대화시키는 최대우도법 $L(\hat{\lambda}, \hat{\theta})$로 추정해야한다 (Jugg ott, 1991).

$L(\hat{\lambda}, \hat{\theta})$를 추정하는 경우는 두 가지 방법이 있는데 첫째로는 θ가 연관 되어있다는 대립가설 (H_1 of linkage : $H_1 : \theta < \frac{1}{2}$) 하에서 $\hat{\lambda}$를 추정하는 것이고 다음으로는 θ가 연관 되어있지 않다는 귀무가설 (H_0 of absence of linkage : $H_0 : \theta = \frac{1}{2}$) 하에서 추정하는 것이다. 이런 경우 Z_{max}는 아래와 같이 나나내어 질 수 있다.

$$Z_2 = \log_{10} \frac{L(\hat{\lambda}, \hat{\theta})}{L(\hat{\lambda}, \theta = \frac{1}{2})} \qquad , \quad H_1 \text{ 하에서}$$

$$Z_1 = \log_{10} \frac{L(\hat{\lambda}, \hat{\theta})}{L(\hat{\lambda}_{\theta=\frac{1}{2}}, \theta=\frac{1}{2})} \quad , \quad H_0 \text{ 하에서}$$

Z_2 은 $\theta = \frac{1}{2}$ 이라는 제약(restriction)조건이 있기 때문에 자유도 1의 값을 가지고 Z_1 는 $\theta = \frac{1}{2}$ 이라는 제약조건외에 $\hat{\lambda}_{\theta=\frac{1}{2}}$ 이라는 조건이 추가되므로 자유도 2를 가지게 된다. 그러므로 Z_1 이 Z_2 보다 연관정도를 측정하는데는 더욱 효율적이라 할 수 있다. 즉, 같은 LOD Score 를 가진다면 Z_1 이 Z_2 보다 더욱 강한 연관정도를 나타낸다고 볼 수 있다.

일반적으로 침투율 λ와 재조합률 θ의 상관도는 λ 추정에 영향을 미친다고 알려져 있는데 부모의 유전형을 알 경우(phase-known) 부모의 유전형을 한쪽은 표식유전자와 질병유전자 모두 동형(double homozygote)으로 가정하고 다른 한쪽은 이 두유전자 모두를 이형(double heterozygote)으로 가정하는 교배방법인 double backcross mating을 고려할 때 λ 와 θ의 상관관계 ρ는 다음과 같이 추정된다.

$$\rho(\hat{\lambda}, \hat{\theta}) = \frac{\lambda(1-2\theta)\sqrt{\theta-\theta^2}}{\sqrt{1-\lambda+2\lambda^2\theta(1-\theta)(1-2\theta+2\theta^2)}}$$

이때 λ와 θ의 범위는 $0 \le \lambda \le 1,\ 0 \le \theta \le \frac{1}{2}$ 이 되는데, $\theta < \frac{1}{2}$ 이면 λ와 θ는 양의 상관관계를 갖고 $\theta = \frac{1}{2}$ 이면 독립이 된다.

이와 같이 λ와 θ의 결합 추정치(joint estimate)는 그 추정치가 가지는 실제적인 범위 때문에 서로간에 연관성을 가지게 되고 이로 인해서 두 모수를 추정할

경우 그 추정치는 대체적으로 강하게 편의(strongly biased)되어 지게 된다. 또한 이런 우도를 사용한 추정은 그 속성이 매우 복잡하기 때문에 실제 자료에서의 추정에는 많은 어려움이 있다.

2. 4. 3 질병이 발병된 사람들을 대상으로 한 연관성 검정

앞절에서 언급했던 결합우도를 사용한 추정은 질병이 있는 사람과 질병이 없는 사람 모두를 대상으로 한다. 그러나 질병을 가지고 있는 사람들은 질병이 없는 사람들에 비해 유전자 재조합에 대한 정보를 많이 가지고 있는데 즉, 그 사람들의 염색체 위에 질병 유전자가 있을 확률이 높다. 이러한 이유로 질병을 가지고 있는 사람들만을 대상으로 유전자 정보를 작성하면 모든 사람을 대상으로 하는 것보다 재조합에 대한 정보의 손실도 크지 않으면서 비용과 시간 면에서도 훨씬 더 효율적일 수 있다.

이 때문에 이 절에서는 주로 질병을 가지고 있는 사람만을 대상으로 질병유전자와 표식유전자 사이의 연관성 존재여부(existence of linkage)에 대한 검정 방법을 보이고자 한다.

연관성 검정에서 가장 많이 쓰이는 방법 중 첫번째는 병에 감염된 친족에 대한 디자인(affected sibpair design)을 이용한 것으로 Penrose (1935)에 의해 처음 제안되었다. 이 방법을 사용하기 위해서는 몇 가지 가정이 선행되어야 하는데, 첫번째로 표식유전자는 완전한 유전자다형성 (full polymorphic)을 만족해야 한다. 즉, 부모의 대립형질은 12×34 와 같이 네 유전자 모두 달라야 한다. 두번째로 질병유전자는 유전되어야 하고 반드시 발병되어야 한다. 하지만 이것은 불완전 침투의 경우 때문에 항상 만족되지 않는다. 검정하고자 하는 가설은 다음과 같다.

H_0 : 표식유전자와 질병유전자는 연관되어 있지 않다($\theta = \frac{1}{2}$).

H_1 : 두 유전자는 연관되어 있다 ($\theta < \frac{1}{2}$).

두 명의 발병된 형제 또는 자매가 공유(share)하고 있는 표식유전자의 수를 S_m 이라 하면 각 S_m 이 나타날 확률은 아래와 같고 일반적으로 부모의 표식유전자의 유전형을 알 경우 공유하는 표식유전자의 수는 이항 분포(binomial distribution)를 따른다.

$$P(S_m = 0) = \frac{1}{4} \quad , \quad P(S_m = 1) = \frac{1}{2} \quad , \quad P(S_m = 2) = \frac{1}{4}$$

위의 S_m 을 사용하여 간단한 연관성 검정을 할 수 있는데 n 명의 형제 또는 자매 (sibpairs)가 있을 경우($n = n_0 + n_1 + n_2$), n_0 , n_1 , n_2 각각 공유하고 있는 표식유전자의 수가 0, 1, 2 라고 가정한다면 다음과 같은 Q 검정통계량 (test statistic)을 사용할 수 있는데 Q 통계량은 자유도 2 인 χ^2 분포를 따른다.

$$Q = \sum_{i=0}^{2} \frac{[n_i - np(S_m = i)]^2}{np(S_m = i)} \sim \chi^2_2$$

그러나 위 식은 표식유전자와 질병유전자가 연관되어 있지 않다는 귀무가설 하에서 성립된 통계량이므로 만약 표식유전자와 질병유전자가 연관되어 있다는 대립가설 하에서는 통계량의 검정력(power of test)이 떨어지게 된다. 즉, 대립가설 하에서는 $S_m = 1$, $S_m = 2$ 가 나타날 확률이 $S_m = 0$ 보다 높아지게 된다. 이 때문에 검정력을 좀더 높일 수 있는 아래와 같은 Z 통계량을 사용할 수

있는데 이 Z 통계량은 표준정규분포(standard normal distribution)를 따른다.

$$Z = \frac{n_2 - n_0}{\sqrt{\frac{n}{2}}} \sim N(0, 1)$$

그러나 위의 병에 감염된 친족에 대한 디자인은 한 가족의 형제, 자매에서만 쓰일 수 있는 디자인이기 때문에 여러 가족에 대해서는 일반화되어 있지 않다.

이러한 단점을 보완한 것으로 Green, Low, Woodrow (Green et al, 1983; Green et al, 1984)는 N-검정을 제안하였는데 이 N-검정은 De Vries (De Vrieset al, 1976)에 의해 만들어진 F-검정과 유사하다.

N-검정을 하기 위해서는 앞에서 언급했던 디자인에서의 가정과 같이 부모의 표식유전자는 완전한 유전자다형성을 만족해야 하고 또, s_i ($s_i \geq 2$)는 i번째 ($i = 1, 2, \ldots, k$) 가족에서 질병에 감염된 형제, 자매라 가정 하고 N_i 를 부모로부터 받은 유전자의 반수체(haploid)중 가장 많이 나타나는 대립유전자의 합이라 정의할때 N_i 의 정의에 의해서 통계량 N 은 다음과 같이 정의된다.

$$N = \sum_{i=1}^{k} N_i$$

이 통계량은 각 가계(family)들이 독립이고 또한 그에 속한 구성원의 수가 많아 질때($k \rightarrow \infty$) 근사적 (approximately)으로 정규분포 $N(\mu, \sigma^2)$를 따른다. 이 통계량을 이용하여 가설의 검정을 위한 Z 검정통계량을 다음과 같이 계산할 수 있다. 우선 N_i 의 평균($E(N_i)$)과 분산($V(N_i)$)을 다음과 같이 계산한다.

$k(m)$: 어머니로부터 감염된 s_i 의 표식유전자의 수

$k(f)$: **아버지로부터 감염된** s_i **의 표식유전자의 수**

$k(m)$, $k(f)$**를 위와 같이 정의하면 표식유전자와 질병유전자가 연관되어 있지 않다는 귀무가설 하에서 아래와 같은 조건들이 성립한다.**

$$k_i(m) \sim B\left(s_i, \frac{1}{2}\right),\ k_i(f) \sim B\left(s_i, \frac{1}{2}\right)$$

$$N_i(m) = \max\left(k_i(m), 1 - k_i(m)\right)$$

$$N_i(f) = \max\left(k_i(f), 1 - k_i(f)\right)$$

이로부터 $N_i(m)$ **의 분포는 다음과 같다.**

$$\Pr[N_i(m) = j] = \begin{cases} 2\binom{s_i}{j}\left(\frac{1}{2}\right)^{s_i} & \text{for all } j > \frac{s_i}{2} \\ \binom{s_i}{j}\left(\frac{1}{2}\right)^{s_i} & \text{for } j = \frac{s_i}{2} \end{cases}$$

$N_i(f)$ **의 분포도 동일하게 구성될 수 있다.**

$N_i(m)$**와** $N_i(f)$**가** iid(independent identically distribution)**를 따를 때** $N_i = N_i(m) + N_i(f)$**로 나타내어지고 이때** $E(N_i)$**와** $V(N_i)$**는 아래와 같이 계산되어진다.**

$$E(N_i) = 2E(N_i(m)),\quad V(N_i) = 2V(N_i(m))$$

이로부터 N**통계량의 평균과 분산을 구할 수 있다.**

$$E(N) = \sum_{i=1}^{k} E[N_i],\qquad V(N) = \sum_{i=1}^{k} V[N_i]$$

위 결과로부터 검정통계량 Z은 아래와 같이 나타내어진다.

,

$$Z = \frac{N - E(N)}{V(N)} \sim N(0, 1) \quad as \quad k \to \infty$$

다음으로 Green 과 Grennan (Green et al, 1991)은 더욱 일반적인 상황에서 사용되어질 수 있는 아래와 같은 T 통계량을 제시하였는데 이 방법은 부모의 표식유전자가 완전한 유전자다형성을 만족하지 않고 부모의 표식유전자의 유전형도 모를 경우를 고려한 방법이다. 이 방법은 모든 자손의 표식유전자를 통해 부모의 유전형을 추정하는 방법인데 1986년 Lange 의해 제안된 방법과 유사하다.

$$Z = \sum_{i=1}^{s} \sum_{j=i+1}^{s} X_{ij}$$

X_{ij} : i 와 j 가 공유 하고 있는 표식유전자의 수.

s : 발병된 형제 또는 자매의 수.

위와 같을 때 T 검정통계량은 다음과 같다.

$$T = \sum_{i=0}^{k} \frac{Z_i}{\sigma_{s_i}^2}$$

$\sigma_{s_i}^2$: k 개의 가계 중 발병된 형제 자매의 최소분산(minimum variance)

2. 4. 4 질병 위험률을 이용한 재조합률의 추정

발병된 자손들만을 대상으로 침투율과 재조합률을 추정하는 것은 전체를 대상으로 하는 것보다는 비용이나 시간 면에서 훨씬 효율적이지만 발병되지 않은 자손들도 질병 유전자를 보유하고 있기 때문에 불완전 침투에서의 침투율을 추정할 경우 과추정(overestimate)될 수 있어서 상당한 편의를 가지게된다. 또한 불확실한 침투율의 추정으로 재조합률도 불확실한 추정량을 가질 수 있다.

그러므로 발병된 사람과 발병되지 않은 사람 모두를 대상으로 침투율과 재조합률을 추정할 수 있는데, 2. 4. 2 절에서 언급하였던 λ 와 θ의 결합 우도로 추정하는 방법이다. 그러나 이 방법은 모수(parameter : λ 와 θ) 2개를 동시에 추정해야 하므로 추정식이 매우 복잡해져서 현실적으로 추정하기가 힘들고 막대한 비용과 시간의 투자를 필요로 한다.

이러한 이유로 1990년에 Risch 는 침투율 λ를 결합우도로 추정하는 대신에 질병의 발병률(onset rates , disease risks)을 통해 추정하고 이에 따라 재조합률 θ를 추정하는 방법을 제안하였는데 이 방법은 모든 자손들을 대상으로 발병률을 추정하고 발병된 자손들만을 대상으로 유전자형을 조사하여 결합우도로 재조합률을 추정하는 방법이다. 이는 결합우도를 사용한 추정과 발병된 사람을 대상으로 한 추정의 모두 고려한 것이다.

발병률의 추정은 다음과 같이 할 수 있다. 발병률에 관한 몇 가지 정의를 하면 아래와 같다.

λ_p : 모집단의 발병률 (disease risk of population).

λ_o : 부모중 한사람이 발병되었을 때 그 자손이 발병될 확률.

λ_s : 형제, 자매중 한 명이 발병되었을 때 다른 형제자매가 발병될 확률.

λ_m :일란성 쌍둥이(monozygotic twin)중 한 명이 발병되었을 때 나머지 한 명이 발병될 확률.

단일 로쿠스(single locus)위에 n 개의 대립유전자 $g_1, g_2, \ldots, g_n$ 가 있다고 가정 할때, g_i 의 모집단 빈도 (population frequency)를 t_i 라하고 유전형 $g_i g_j$ 의 침투율을 f_{ij} 라 가정한다. 또 첫번째 사람이 발병되었으면 X_1 은 1의 값을 가지고 발병되지 않았으면 0 의 값을 가진다고 한다. 2번째 사람 X_2 도 앞의 경우와 같다. 이런경우 대립유전자의 빈도는 HWE (Hardy - Weiburg Equilibrium)를 따른다고 가정하면 모집단의 유병률(prevalence) K 는 아래와 같이 정의된다.

$$K = \sum_{i=1}^{n} \sum_{=1}^{n} t_i t_j f_{ij}$$

또한, 관계(두 사람의 관계, s : sibling, o : offspring, m : MZ twin)에 따라 질병유전자를 공유할 확률 K_R 은 다음과 같이 정의된다.

$$K_R = E(X_2 \mid X_1 = 1) = K + \left(\frac{1}{K}\right) = Cov(X_1, X_2) + K^2$$

$Cov(X_1, X_2)$: X_1 과 X_2 사이의 공분산

따라서 관계에 따른 발병률 λ_R 은 다음과 같이 나타낼 수 있다.

$$\lambda_R = \frac{K_R}{K} = 1 + \left(\frac{1}{K^2}\right) Cov(X_1, X_2)$$

위 관계에서 인구집단 조사(population survey)를 통해 λ_p , λ_o, λ_s 는 비교적 쉽고 정확하게 추정할 수 있으나 λ_m 은 드문 현상이기 때문에 추정하기가 쉽지 않다. 그러므로 λ_p , λ_o, λ_s 통해 다음과 같이 λ_m을 추정 할 수 있다 (Risch, 1990).

$E(X_1) = E(X_2) = \lambda_p$ 이고

$Cov(X_1, X_2) = E[(X_1 - \lambda_p)(X_2 - \lambda_p)]$

$= E(X_1 X_2) - \lambda_p^2 = p(X_1 = X_2 = 1) - \lambda_p^2$

만약 두 사람의 상관관계가 없다면 공분산(covariance)은 0 이 된다. 위 공분산을 ibd(identity by descent ; 같은 조상으로부터 물려받은 한 쌍의 대립유전자)와 조건부 확률로 다시 표현하면 아래와 같다.

$$Cov(X_1, X_2) = \sum_{i=0}^{2} E[(X_1 - \lambda_p)(X_2 - \lambda_P) \mid ibd = i] P(ibd = i)$$

$$= \sum_{i=1}^{2} E[(X_1 - \lambda_p)(X_2 - \lambda_P) \mid ibd = i] P(ibd = i)$$

$V_1 = E[(X_1 - \lambda_p)(X_2 - \lambda_P) \mid ibd = 1]$

$V_2 = E[(X_1 - \lambda_p)(X_2 - \lambda_P) \mid ibd = 2]]$

이라 하면 $Cov(X_1, X_2) = V_1 P(ibd = 1) + V_2 P(ibd = 2)$ 이다.

다른 표현으로 하면 다음과 같다.

$Cov(X_1, X_2) = p(X_1 = X_2 = 1) - \lambda_p^2$

$= \Pr(Both\ X_1 \text{ and } X_2\ are\ affected) - \lambda_p^2$

$= \Pr(X_1\ is\ affected \mid X_2\ is\ affected)\Pr(X_2\ is\ affected) - \lambda_p^2$

$= \lambda_R \lambda_p - \lambda_p^2$

그러므로 $V_1 P(ibd = 1) + V_2 P(ibd = 2) = \lambda_R \lambda_p - \lambda_p^2$ **이다.**

위에서 아래와 같은 세 방정식을 얻을 수 있는데 다음과 같다.

$$V_2 = \lambda_m \lambda_p - \lambda_p^2$$

$$V_1 = \lambda_o \lambda_p - \lambda_p^2$$

$$V_1/2 + V_2/4 = \lambda_s \lambda_p - \lambda_p^2$$

그러므로 다음과 같이 λ_p, λ_o, λ_s **통해** λ_m**를 추정할 수 있는 식이 정의된다.**

$$\lambda_m = 4\lambda_s - 2\lambda_o - \lambda_p$$

다음으로 우도식을 사용하여 재조합률 θ **를 추정하면 아래와 같다.**

p_i**을 동일한 대립 유전자** i**를 가지는 형제, 자매가 발병될 확률이라고 정의하고** S_m $(= n_i)$**을 2. 4. 3절과 같이 정의한다면 다음과 같은 식으로 표현할 수 있다.**

$$L(\theta) \propto \prod_{i=0}^{2} p_i^{n_i}$$

$$p_i = P(S_m = i) = \frac{p(\text{두명의 형제 모두 발병} \mid S_m = i)p_o(S_m = i)}{P(\text{두명의 형제 모두 발병})}$$

보통 ibd를 두 사람간의 ibd 한 대립유전자의 수 와 같은 개념으로 정의하는데 만약 두 사람의 최근 조상이 다르다면 한 쌍의 대립유전자의 ibd(=π)는 0 이라고 볼 수 있다. 일반적으로 부모와 자녀, 두 명의 형제 또는 자매의 ibd의 평균은 같고 분포는 같지 않다. 그러므로 ibd는 대립유전자들의 분포에 독립이다.

π_m , π_t 를 아래와 같이 정의하고 표식유전자는 유전자 다형성을 만족한다고 가정한다면 위 식 p_i 의 분자부분은 [표 2-1]에 의하여 다음과 같이 계산되어 진다.

$$p(\text{두명의 형제 모두 발병} \mid S_m = \pi_m = i)$$

$$= \sum_{j=0}^{2} p(\text{두명의 형제 모두 발병} \mid \pi_t = j)\, p(\pi_t = j \mid \pi_m = i)$$

π_m : 두명의 형제또는 자매에 의해 공유하고 있는 ibd 에 의한 표식유전자의 수. 즉, $S_m = \pi_m$

π_t : 두명의 형제또는 자매에 의해 공유하고 있는 ibd 에 의한 질병 유전자의 수.

[표 2-1] $p(\pi_t = j \mid \pi_m = i)$ 의 값 (Haseman et al, 1972)

π_m \ π_t	2	1	0
2	ψ^2	$2\psi(1-\psi)$	$(1-\psi)^2$
1	$\psi(1-\psi)$	$\psi^2 + (1-\psi)^2$	$\psi(1-\psi)$
0	$(1-\psi)^2$	$2\psi(1-\psi)$	ψ^2

$\psi = \theta^2 + (1-\theta)^2$

2. 4. 5 불완전 침투의 효율성

불완전침투는 유전형에 의해 표현형을 정확히 알 수 없기 때문에 재조합률을 추정할 경우 완전침투보다는 더 많은 정보의 손실(loss of information)이 있다 (Jurg ott 1991). 즉, 재조합률 추정의 효율성(efficiency)이 떨어지게 된다.

일반적으로 하나의 표식유전자를 가지고 질병유전자를 추정하는 두 유전자 거리 분석(two-point analysis)의 경우 침투의 효율성을 평가하는 기준으로 ELOD(expected lod score)를 주로 사용하는데, ELOD 는 LOD Score 의 가중평균(weight average)으로 다음과 같이 정의된다.

$$ELOD = \sum_x P(x:r)\, Z_x(\theta)$$

$$Z(\theta) = LOD = \log_{10} \frac{L(\theta)}{L(\theta = \frac{1}{2})}$$

r : 실 재조합률(true recombination fraction)

$P(x:r)$: 가중치 $[\sum_g P(x \mid g) P(g:r)]$

x : 표현형

g : 유전형

θ : 재조합률

또, 침투의 효율성을 평가할때 주로 사용하는 교배 방법(mating type)은 부모의 유전형을 한쪽은 표식유전자와 질병유전자 모두 동형(double homozygote)으로 가정하고 다른 한쪽은 이 두유전자 모두를 이형(double heterozygote)

으로 가정하는 방법(double backcross mating)이 가장 많이 쓰이는데, 이 방법은 다른 교배방법에 비해 보다 많은 재조합률에 관한 정보를 제공하기 때문이다. 이 때문에 침투의 효율성은 어떤 교배 방법을 선택하느냐에 따라 달라질 수 있다.

침투에서의 효율성은 다음과 같이 완전침투의 ELOD에 대한 불완전 침투의 ELOD 의 비율(ratio)로 나타내어진다.

$$R = \frac{ELOD(\ incomplete\ \ penetrance\ \lambda\ )}{ELOD(\ full\ \ penetrance\ \lambda = 1\ )}$$

또한 ELOD의 상대적 비율인 R은 재조합률 보다는 침투율에 따라 더 큰 차이를 보인다 (Elandt -Johnson, 1971).

2.4.6 나이에 따라 변화하는 침투

가장 대표적인 불완전 침투 중 하나가 나이에 따라 침투율이 변하는 경우이다 (Jurg ott, 1991).

이러한 나이에 따라 변화하는 침투(age-dependent penetrance)를 보는 관점은 세 가지가 있는데, 첫번째로 가계도에서 보균자를 나이에 따라 k 개의 집단으로 층화시켜 보는 방법이다. 각 k 개의 그룹에서 침투율 λ를 구하는데, 이때 k번째 집단은 은 k-1 번째와 k 번째에 있는 보균자들로 구성되고 침투율 λ 는 다음과 같이 추정된다.

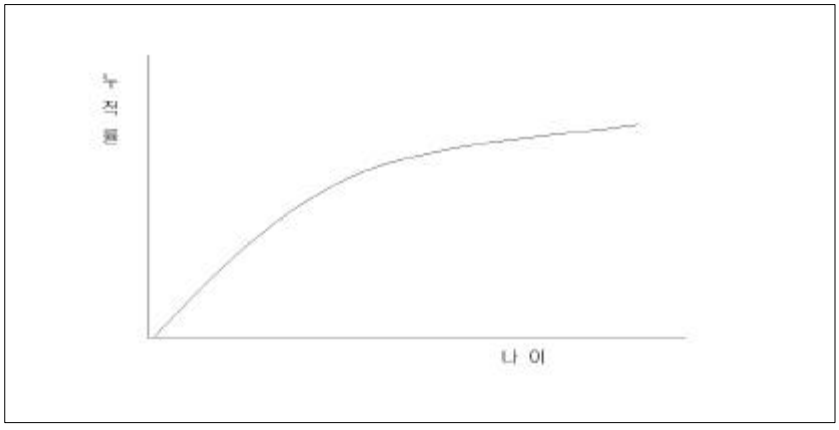
$$\hat{\lambda_k} = \frac{k-1 \text{ 과 } k \text{ 사이에서 발병된 사람의 수}}{\text{총 보균자의 수}}$$

이 방법은 침투율 추정의 일반적인 방법이지만 모집단에서 보균자를 찾는 것이 쉽지 않고 또한 발병된 사람도 그 병이 언제 발병이 되었는지를 판단하기 어렵기 때문에 발병된 사람의 나이 집단(age class)을 알기 어렵다는 단점이 있다.

두 번째로 나이에 따른 발병(onset)의 누적분포(cumulative distribution)를 이용하는 발병나이 곡선(age-of-onset curves)방법 이다. x_k 를 k 번째 그룹의 상한선(upper boundary)이라 하면 x_k 의 침투율은 k 번째 나이 그룹까지 발병된 사람들의 총 수가 된다. 이로부터 누적률(cumulative proportion) $F(x_k)$를 구할 수 있는데 다음과 같다.

$$F(x_k) = \frac{k\text{번째 그룹까지 발병된 사람의 수}}{\text{총 발병자 수}}$$

나이에 따른 누적분포 곡선은 대체적으로 [그림 2-1]와 같은 형태를 보인다.



[그림 2-1] 나이에 따른 누적분포 곡선

세 번째로 각 나이의 그룹별로 침투율을 계산하는 대신에 발병-나이 곡선의 $F(x)$를 함수 형식으로 추정하는 방법이다. 즉, 나이를 구간별로 보지 않고 하나의 연속형 변수(continuous variable)로 보는 것이다. 이 방법에는 크게 두 가지 접근방법이 있는데 로지스틱 분포(logistic distribution)로 보는 방법과 균일 분포(uniform distribution)로 보는 방법이다.

우선 로지스틱 분포로 접근할때 $F(x)$ 는 다음과 같은 분포를 보인다.

$$F(x) = \frac{1}{1 + \exp(-1.8138 Z_i)}$$

여기에서 $Z_i = \frac{(x_i - \mu)}{\sigma}$ 이고 μ 와 σ 는 발병된 나이의 평균과 표준편차이다.

다음으로 균일 분포로 접근 할때 $F(x)$ 는 아래와 같은 분포로 정의될 수 있다.

$$F(x) = \begin{cases} 0 & \text{if } x \le x_1 \\ (x - x_1) / (x_2 - x_1) & \text{if } x_1 < x < x_2 \\ 1 & \text{if } x_2 \le x \end{cases}$$

x_1 : 발병된 사람들의 그룹 중 처음 병이 나타난 나이.

x_2 : 발병된 사람들의 그룹 중 가장 늦게 병이 나타난 나이.

2. 5 가계도 자료의 정보 충분성의 측정

2. 5. 1 가계도 자료의 정보량 측정

연관성 검정을 하기 위하여 가계도를 모았을 때 그 가계도가 과연 이 분석에 얼마나 효율적인가? 다시 말해서 이 가계도가 가지고 있는 자료로 알고 싶은 특성유전자 즉, 질병유전자를 얼마나 잘 찾아낼 수 있는가 하는 것은 중요한 문제이다.

이런 가계도 자료의 정보 충분성을 검정하는 유형은 일반적으로 두 가지 알려져 있는데 첫번째 유형으로는 LOD score를 이용하여 검정하는 것이고 두번째로는 Fisher's Expected information (Fisher, 1925)을 이용하는 방법이다.

첫 번째 유형으로서 먼저 ELOD를 들 수 있는데, ELOD는 모든 발생가능한 결과(outcome) 즉, 각 재조합률 θ 에서 유전자 재조합이 발생한 명수에 대한 LOD Score의 가중평균으로 앞에서 정의한 바 있는데, 여기에서 가중치(weight)는 발생가능한 각 결과에서의 유전자 재조합이 발생할 확률로 나타낼 수 있다.

i번 째 발생결과 에서의 LOD Score 를 $Z_i(\theta) = \log_{10}[\frac{L(\theta)}{L(\theta=\frac{1}{2})}]$

이라 하고 가중치를 $P(i)$ 라 하면 ELOD 는 다음과 같이 나타내어진다.

$$ELOD = \sum_i P(i) Z_i(\theta)$$

ELOD 는 실 재조합률 r 에 따라 달라지는데 이것은 가중치 $P(i)$ 가 실 재조합률의 영향을 받기 때문이다. 또 재조합 확률 θ와 실 재조합률 r이 같

을때 ELOD 는 최대 값을 가지게 된다 (Jurg ott,1991).

ELOD 와 함께 정보의 충분성을 판단하는 것으로 최대 LOD Score 가 특정 상수 보다 클 확률 $P(Z_{max} \geq c)$ 이 있는데, 일반적으로 상수 c 는 3 으로 비교하는데, 상황에 따라 더 낮은 값을 쓰기도 한다.

또, $P(Z_{max} \geq c)$의 방법은 주어진 가계도 자료로 연관성 분석을 할때 표식유전자와 질병유전간의 연관성을 바르게 판단할 확률서 그 자체로 직접적 해석이 가능하다. 그러나 ELOD는 하나의 가계도에서 계산된 ELOD 값으로만 평가하는 것이 아니라 구조가 다른 가계도를 서로 비교함으로서 상대적인 정보의 충분성을 평가하기 때문에 다른 가계도에 대해서 상대적으로 해석이 된다.

위에 언급한 ELOD 와 $P(Z_{max} \geq c)$ 방법 이외도 정보의 충분성을 평가하는 한가지 방법이 더 있는데, 최대 Lod Score의 기대값인 $E(Z_{max})$를 구하는 것이다. ELOD 와 비슷한 개념으로서 ELOD 보다는 항상 큰 값을 가지게 된다.

두 번째로, Fisher's expected information도 정보의 충분성을 검정하기 위한 방법인데, 이 방법은 아래와 같이 전체 정보의 양을 추정함으로서 가계도 자료의 정보 충분성을 판단한다.

$$\hat{I}(\hat{\theta}) = - \frac{d^2 \ln L(\theta)}{d\theta^2}$$

또, $i(\hat{\theta})$ 을 하나의 가계도 또는 한 사람이 주는 정보의 양이라 하고 전체 n 개의 가계도 또는 n 명이 있다고 가정한다면 Fisher's information 에서는 다음 식이 성립한다.

$$I(\hat{\theta}) = n\,i(\hat{\theta})$$

본 연구에서는 상기의 여러 방법 중 실제 상황에서의 간편성을 고려하여 $E(Z_{\max})$와 $P(Z_{\max} \geq c)$를 중심으로 심혈관질환(cardiovascular disease) 자료에 적용하였다.

2.5.2 연관성분석에서 검정력

간단한 유전학적 구조를 가질때 가계도의 정보 충분성은 윗 절에서 언급한것과 같이 이론적으로 직접 계산되어질 수 있다. 그러나 치매의 종류인 알쯔하이머(Alzheimer disease)나 심혈관질환같은 대분분의 질병들은 간단학 유전학적 구조보다는 여러 유전자가 복합적으로 작용하는 복잡한 유전학적 구조를 가지거나 불완전침투의 양상을 보이고 있기 때문에 직접적인 계산보다는 모의 실험(simulation)을 통하여 가계도의 검정력을 추정하는 것이 현실적으로 좀더 간단하다 (Thompson et al. 1978; Skolnick et al. 1984).

연관성 분석을 할 경우에는 질병과 연관되어 있는 표식유전자의 유전자형이나 질병유전자의 유전자형 같은 여러 가지의 가계도에 대한 충분한 정보(sufficiency of information)가 있어야 하므로 연관성 분석시 충분한 정보를 가지고 있는 가계도를 수집하는 것은 매우 중요한 과정이다. 그러나 이런 정보 수집에는 현실적으로 여러 가지 많은 어려움이 따른다.

이러한 이유 때문에 연관성분석전 가계도의 검정력을 추정하는 것인데, 가계도 검정력(power of pedigree)의 추정은 가계도 자료의 정보의 충분성 여부를 판단함으로서 가계도의 여러 자료들을 수집할 때 도움을 준다 (Ploughman and Boehnke, 1989).

이 방법은 일반적으로 표식유전자의 유전형 조사전에 시행하는 방법으로서 필요한 가계도의 수나 질병유전자와 관련이 있는 적절한 표식 유전자의 정보등을 제공함으로서 분석 전 좀더 효율적인 연구 디자인 설계를 할 수 있다.

1986년 Boehnke는 멘델리안 법칙을 따르는 완전 침투의 경우를 가정하고 가계도의 정보충분성 여부를 판단하는 방법을 제시하였는데, 이 방법은 개개인의 질병의 유전자형이 모두 조사되었을 경우만 가능했었다 (Jurg ott, 1991). 그 이후에 불완전 침투이거나 개인의 질병 유전자형을 알 수 없는 경우 등 더욱 일반적인 상황에 적용할 수 있는 방법들이 제시되었다 (Ploughman and Boehnke, 1989).

제 3 장 모의실험 기법과 모형

3. 1 모의 실험을 통한 추정

모의실험을 통한 가계도 자료의 검정력을 추정하기 위해서는 몇 가지 가정이 필요한데, 첫째로 각 개인의 질병의 표현형, 즉, 질병 발병의 여부를 알아야 하고 둘째로는 표식유전자와 질병유전자는 연관되어 있어야 하고 세째로는 가계도의 구조는 분석시 적합한 구조라고 가정한다 (Ploughman and Boehnke, 1989). 이러한 가정을 통해 가계도 자료의 검정력을 다음 단계를 통해 추정할 수 있다.

첫번째 단계로 알고있는 질병의 유전형과 표현형의 정보를 이용하여 가계도안의 각 개인에 대한 질병 유전형의 확률 즉, 질병의 유전형과 표현형이 주어졌을 때 개인의 질병유전형의 조건부 확률 (coonditional probability)을 계산한다.

질병의 유전형의 벡터를 $g = (g_1, \ldots, g_n)$ 라하고 표현형의 벡터를 $x = (x_1, \ldots, x_n)$ 이라 할 때, 질병 유전형의 조건부 확률은 $P(g \mid x)$로 다음과 같이 표현할 수 있다.

n 명으로 구성된 가계도에서 1, 2, 는 부모라 하고 자손을 3, … , n 이라 하면 질병의 유전형의 조건부 확률은 다음과 같이 나타낼 수 있다.

$$P(g \mid x) = \prod_{k=1}^{n} p(g_k \mid g_1, \ldots, g_{k-1}, x)$$

$$= P(g_1 \mid x) P(g_2 \mid g_1, x_2, \ldots, x_n) \prod_{k=3}^{n} P(g_k \mid g_1, g_2, x_k)$$

이 조건부확률은 Elston-Stewart 연산(algorithm) (Elston and Stewart, 1971)을 일반화시킨 Lange 와 Boehnke(1983)의 연산을 이용한 우도식을 사용하여 계

산되어 지는데, 이 우도식은 무작위 교배(random mating)를 가정하였고 환경적인 효과(environment effect)로 인한 유전자의 공유가 없다고 가정하였다. 이러한 가정이 있을때 우도식은 아래와 같이 표현되어질 수 있다.

$$L(x) = \sum_{g_1} \cdots \sum_{g_n} \prod_i P(x_i \mid g_i) \prod_j P(g_j) \prod_k P(g_k \mid g_l, g_m)$$

$P(x_i \mid g_i)$: i번째 사람의 조건부 확률

(x_i : i번째 사람의 표현형 , g_i : i번째 사람의 유전형)

$P(g_j)$: 부모가 가계도 안에 없는 j 번째 사람의 유전형의 사전확률

$P(g_k \mid g_l, g_m)$: 자손 k 의 부모 l, m의 유전형이 주어졌을때

k의 유전형 g_k 의 조건부 확률.

두번째 단계는 첫번째 단계에서 계산된 유전형의 조건부 확률을 이용하여 반복(replicates)된 가계도에서 각 개인의 질병의 유전형을 모의실험 한다.

세번째로 표식유전자와 질병유전자간의 여러 실 재조합률을 가정하고 그 가정된 각 실 재조합률 에서 표식유전자의 유전형을 모의실험 한다.

이때, 가계도안에서 표식유전자 유전형의 대립형질들은 대립형질의 모집단 빈도(population frequency)를 기초로 하여 할당되는데 각 대립형질의 빈도는 HWE (Hardy-Weinberg equilibrium)를 가정한다.

네번째는 전 단계에서 정의된 각 실 재조합확률 에서 가계도의 $Z_{\max}$ (maximum lod score)를 계산한다.

마지막으로 $Z_{\max}$ 분포의 기초통계량을 이용하여 가계도 정보 충분성에 대한 두 가지 정보를 측정할 수 있는데, 하나는 $E(Z_{\max})$ (Expected maximum lod score) 이고 다른 한가지는 $Z_{\max} \geq c$ (c : 특정상수)의 확률이다. 이를 가계도

의 검정력 이라 하는데 즉, 표식유전자와 질병유전자가 연관되어 있다고 할 경우 주어진 가계도의 자료로서 이것을 맞게 판단할 확률이다. 일반으로 상염색체(autosome)일 경우 $Z_{max} \geq 3$이면 연관되어 있다고 보고 성염색체(X-chromosome)일 경우는 $Z_{max} \geq 2$ 이면 표식유전자와 질병유전자가 연관되어있다고 판단하는데 경우에 따라서 상수의 기준은 더 낮아 질 수 있다(Jurg ott , 1991).

또, 첫번째 $E(Z_{max})$는 Z_{max} 의 표본 평균에 의해 추정되고 두번째 $Z_{max} \geq c$ 의 확률은 Z_{max} 의 표본비율 sample proportion에 의해 추정되어진다.

3. 2 모형의 개요

3. 2. 1 모의실험 프로그램의 구성

앞 절에서 언급되었던 가계도의 검정력을 추정하기 위하여 SIMLINK (Version 4.12, 1997)프로그램이 사용되는데, 이는 Boehnke(1986)와 Ploughman, Boehnke(1989)이 제안한 이론을 바탕으로 하여 만들어진 프로그램으로 FORTRAN에 의해 컴파일(compile)되어 MS-DOS 상에서 실행이 된다. 이 프로그램은 가계도의 검정력을 추정하기 위해 크게 세 단계를 거치는데 첫번째로는 질병유전자와 표식유전자의 유전형을 시뮬레이션하고 다음 단계로는 Lod Score나 Location Score (여러개의 표식유전자를 사용할 경우 사용됨)를 계산하고 마지막으로 이 정보들을 이용해 가계도의 정보 충분성 여부를 검정한다. 여기에서 Lod Score 나 Location Score 의 계산은 1988년도에 만들어진 프로그램인 MENDEL(Lange et al., 1988)을 기반으로 하고있다.

또한, 이 프로그램은 가계도의 검정력을 추정 할 때 몇 가지의 가정을 필

요로 하는데 다음과 같다.

첫번째로 질병유전자는 두개의 대립형질로 구성되어야 하고 두 번째로는 가계도에서 각 구성원들의 관계(relationship)와 그들의 질병 표현형을 알아야 하고, 세번째로 질병유전자의 유전형태(mode of inheritance)를 알아야한다. 네번째로는 표식유전자 유전형의 각 대립형질의 빈도는 HWE를 따른 다고 가정하고 다섯번째로 여러 표식유전자를 고려할 경우 표식유전자 사이의 유전적 거리(map distance)는 Haldane(1919)이 제안한 사상함수(mapping function)에 의해서 모의실험 된다고 가정한다. 마지막으로 조사된 가계도에는 일란성 쌍둥이(MZ-twins)가 없다고 가정한다.

3. 2. 2 프로그램에서의 입력과 출력자료

3. 2. 2. 1 입력자료

SIMLINK 에서의 입력 자료는 크게 제어 파일(control file), 로쿠스 파일(locus file), 가계도 파일(pedigree file)로 구성되는데 각 파일은 다음과 같이 구성된다.

첫째로 제어 파일에는 검정력 추정을 위한 일반적인 정보를 기술하게 되는데 우선 첫 번째 열(row) 에는 다음과 같은 제어 정보(control information)를 기술하게 된다.

모의실험의 반복 횟수, 사용되는 표식유전자의 수, 질병유전자의 침투함수(penetrance function)의 양상 즉, 질병유전자의 표현형이 질적인(qualitative)인가 양적인가(quantitative)를 판단하고 질적인 양상을 보이면 선형 함수인가 누적 정규 함수인가를 결정한다. 또, 질병유전자와 표식유전자사이의 실재조합률(true recombination fraction)의 수와 질병유전자의 수에 따른 동질성

(homogeneity)과 이형성(heterogeneity) 여부 등을 기술한다.

다음으로 두 번째 열에는 실 재조합률 이나 유전적 거리에 관한 정보를 기술하게 되는데, 단일 표식유전자를 사용하는 경우에는 실 재조합률을, 두개의 표식유전자를 사용하는 경우에는 두 표식유전자 사이의 유전적거리를 언급해 준다. 그러므로 전자의 경우에는 결정된 각 실 재조합률 하에서 Lod Score를 계산하고 후자의 경우는 유전적거리를 이용하여 Location Score를 계산하게 된다.

세 번째 열부터는 질병유전자 유전형의 대립형질 각 조합에서의 발병되는 최대, 최소 나이와 이때의 침투율 같은 질병유전자의 침투 함수에 따른 정보를 보여주게 된다.

마지막으로 성별의 표시방법과 질병유전자 로쿠스(locus)의 이름, 로쿠스 파일명, 가계도 파일명등을 기술해 주어야 한다.

두 번째로 로쿠스 파일에는 검정력 추정시 사용되는 유전적 정보가 있어야 하는데, 우선 첫 번째 열에는 질병유전자의 다음과 같은 유전적 정보가 언급된다. 질병유전자의 로쿠스의 이름, 이 것은 제어 파일에서 언급되었던 이름과 동일해야 한다. 또, 질병유전자가 상염색체에 위치하는가 성염색체 위에 위치하는가의 여부와 각 대립형질의 수, 대립형질에 따른 질병유전자의 표현형의 수 를 알아야 한다.

두 번째 열부터는 첫 번째 열에서 정의되었던 질병유전자의 각 대립형질의 이름과 모집단 빈도를 언급해 주어야 하고, 이 대립형질에 따라서 나타날 수 있는 질병유전자의 유전형을 기술해 주어야한다. 이때 질병유전자의 대립형질 중 질병을 일으키는 대립유전자가 무엇인지와 질병을 일으키는 대립형질이 우성형인지 열성형인지를 고려해야한다.

표식유전자의 유전적 정보도 위의 경우와 같이 표식유전자의 이름 , 표식유전자의 위치(상염색체 또는 성염색체), 각 표식유전자의 대립형질의 수 등을 알아야 한다.

마지막으로 가계도 파일에서는 실제 검정력 추정에서 사용되는 가계도에 대한 정보가 포함되어 있는데, 여러 개의 가계도를 사용할 경우 순차적으로 가계도 별로 가계도의 이름과 구성원 수를 언급해 주고 그 다음 열에는 각 구성원의 ID, 부모의 ID, 제어 파일에서 언급해 주었던 성별 표시 방법에 따른 성별구분, 구성원의 질병 감염여부, 또는 질병유전자의 보유여부, 불완전 침투일 경우 나이 같은 침투 함수의 변수 등을 기술해 준다.

3.2.2.2 출력자료

제어파일, 로쿠스파일, 가계도 파일에 입력된 정보들을 가지고 가계도의 정보충분성에 대한 검정력을 추정하게 되는데, 크게 여섯가지 결과로 나누어 볼 수 있다.

첫 번째 결과에는 모의 실험에 사용된 정보들을 요약해 언급해 주는데, 질병유전자 로쿠스(locus)의 이름, 모의실험 반복횟수, 실 재조합률, 시험재조합률(test recombination), 유전적 거리등을 보여준다.

두 번째 결과는 가정해준 실 재조합률(또는 유전적 거리)에서 가계도와 각 가계도 합의 최대 Lod Score(또는 Location Score)값의 평균(mean)과 표준 오차(standard error), 최대값을 보여준다.

세 번째 결과에서는 실 재조합확률 에서 각 가계도와 그들의 합의 최대 Lod Score(또는 Location Score)가 특정상수보다 클 확률 즉, $P(Z_{\max} \geq c)$ 값을 보여준다. 이때 특정상수 c 의 값으로는 0.5, 1.0, 1.5, 2.0, 2.5, 3.0 의 값이 사용된다.

네 번째 결과로는 질병유전자가 두개의 표식유전자 사이에 있을 평균확률을 추정해주는데 이 결과는 단일 표식유전자(single genetic marker)를 사용

하거나 또는 하나의 질병 유전자를 가정했을 경우에는 생략된다.

다섯 번째 결과는 같은 염색체 상에 존재하며 함께 유전되는 경향이 있는 유전자 집단에 포함되지 않은 비연관 표식유전자(unlinked marker)를 가정하고 각 시험 재조합률에서 Lod Score(또는 Location Score)의 평균, 표준오차, 최소값, 최대값을 추정해 준다. 여기에서는 비연관 표식유전자가 있을 영역의 추정에 초점을 맞추는데 보통 Lod Score 값이 -2.0 보다 작으면 비연관 영역에 있다고 추정한다(Morton , 1995).

마지막 결과도 비연관 표식유전자를 가정하고 Lod Score(또는 Location Score)가 특정 상수 보다 클 확률을 추정하는데, 여기에서의 특정상수는 -2.0, -1.0 ,0, 0.5, 1.0 ,1.5, 2.0, 2.5, 3.0 의 값이 사용된다.

제 4 장 심혈관질환 자료를 이용한 정보 충분성 측정

4. 1 심혈관질환의 정의및 특성

4. 1. 1 심혈관 질환의 정의

심혈관 질환(cardiovascular disease : CVD) 은 심장질환과 혈관질환으로 구분되는데, 대표적인 심혈관질환으로는 고혈압, 협심증, 심근경색증, 뇌졸증(중풍), 뇌혈관질환 등이 있다.

심장질환 중에 중요한 부분을 차지하는 관상동맥질환은 대개 동맥경화에 의해서 심장에 혈액을 공급하는 관상동맥이 막히거나, 좁아져서 발생하는데 심근경색증이나 협심증이 이에 해당한다. 관상동맥질환은 우리나라에서 지난 30 여년 동안 급격히 증가하여 왔다. 그러나 아직도 미국이나 유럽에 비해서 그 발생률은 $\frac{1}{4}$ 미만으로 낮은 것으로 집계되고 있다. 하지만 미국이나 서구에서는 1962 년을 기점으로 하여 점차로 관상동맥질환에 의한 사망률이 줄어들고 있는 추세이다. 한편 관상동맥질환이 왜 발생하는지에 대해서는 이미 그 원인이 잘 알려져 있는데, 주요한 원인으로는 고지혈증 (고콜레스테롤혈증), 고혈압, 흡연, 당뇨가 있으며, 기타 관상동맥질환의 가족력, 비만, 운동부족, 스트레스 및 여성의 경우 폐경기 이후 등이 위험요인으로 작용하고 이런 여러 가지 위험요인을 복합적으로 가지고 있을수록 발병의 위험도 증가한다. (보건복지부, 1998)

또한 심혈관질환은 나이에 따라 발병률이 변화하는 대표적인 질환으로서 나이에 따라 변화하는 침투 (age-dependent penetrance)의 양상을 보인다.

4. 1. 2 심혈관질환의 유병률과 사망률

심혈관질환은 1970 년대부터 주요한 사망원인으로 등장하여 최근 약간 감소 추세를 보이기는 하나 여전히 사망원인 1순위를 차지하고 있는데, 1999년 심혈관 질환의 사망률은 인구 10만명당 122명으로 남자가 119.8 명, 여자가 124.1 명이다. 각 질환별 사망률로는 고혈압성 질환이 인구 10만명당 7.6 명 (남자 6.3 명, 여자 8.8 명), 허혈성 심장질환이 18.5명 (남자 21.3 명, 여자 15.6 명), 뇌혈관질환이 72.9 명 (남자 69.3 명, 여자 76.7 명) 이다(통계청 , 1999 ; 박훈기 1999).

최근 10년간 뇌혈관질환, 심장질환 등 순환기계 질환의 사망률의 변화 양상을 보면 1990년 인구 10 만명당 163.9 명에서 1992 년 122.0 명으로 25.6% 감소 하였다. 1990 년에 비하여 고혈압성 질환, 동맥경화증의 사망률은 크게 감소하였으며 허혈성 심장질환의 사망률은 1990년 10.4 명에서 1999년 18.5 명으로 77.9 % 증가하였다 (통계청, 1999).

이러한 심혈관 질환은 나이에 따라 유병률이 증가하는 경향을 보이는데 연령별 유병률은[표 3-1] 같다.

[표 4-1] 연령별 심혈관질환의 유병률 (단위 % ; 한국보건사회연구원, 보건사회복지부, 1998)

연령 / 병명	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70이상
고혈압		0.04	0.10	0.77	2.55	6.67	10.81	13.18
중풍 뇌혈관질환	0.02		0.01	0.11	0.23	1.36	2.97	5.30
협심증 심근경색증			0.01	0.04	0.30	1.22	2.09	1.58

4. 1. 3 심혈관질환을 일으키는 유전자

심혈관 질환을 일으키는 질병유전자는 여러 종류가 있지만 이중 19번 염색체 위에 있는 APOE(Apolipoprotein E) 유전자는 LDL(low-density) 콜레스테롤 수치를 상승시켜 심혈관질환을 일으키는 대표적인 유전자로 알려져 있다 (Carlos Lahoz et al , 2001). 이 APOE 유전자는 $\varepsilon 2$, $\varepsilon 3$, $\varepsilon 4$ 의 3가지 대립유전자로 구성되어 있는데 이중 $\varepsilon 2$ 와 $\varepsilon 4$ 가 우성형질의 질병유전자로 알려져 있다.

APOE 의 각 대립형질의 모집단 빈도는 인종에 따라 약간 씩 다르지만 대규모의 Framingham 연구에 따르면 $\varepsilon 3$:0.802, $\varepsilon 4$:0.119, $\varepsilon 2$:0.079 으로 나타난다 (Carlos Lahoz et al , 2001). 또 유전자가 $\varepsilon 4/\varepsilon 4$ 의 대립형질로 구성되어 있을 때는 90 %이상의 발병률을 보이고 하나의 대립형질만 $\varepsilon 4$ 이고 다른 하나는 $\varepsilon 2$ 또는 $\varepsilon 3$ 로 구성되어 있을 때는 약 50% 정도의 발병률을 보인다.

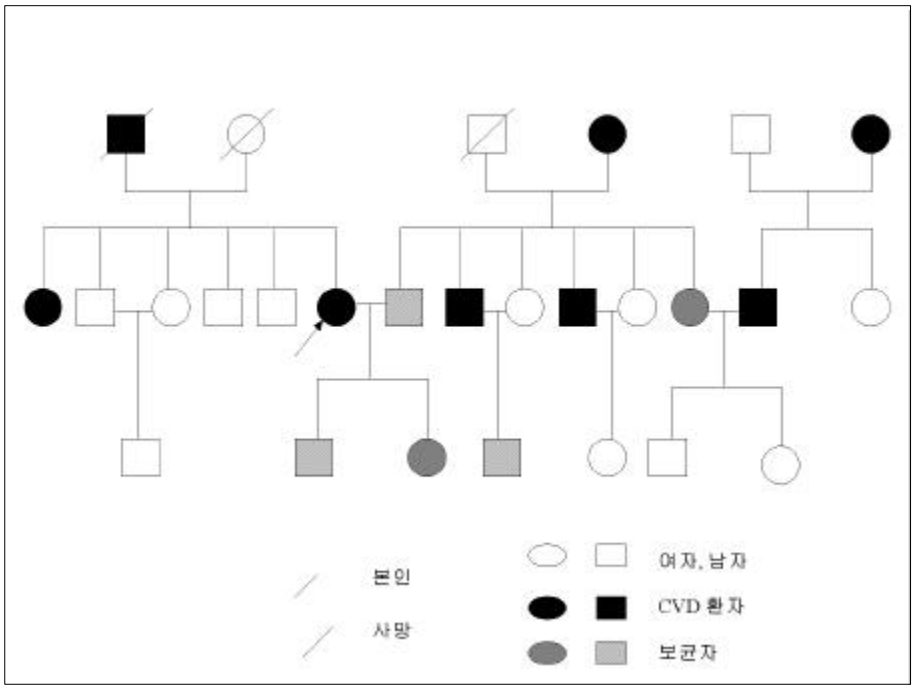
4. 2 심혈관질환 자료에 대한 설명

4. 2. 1 자료현황

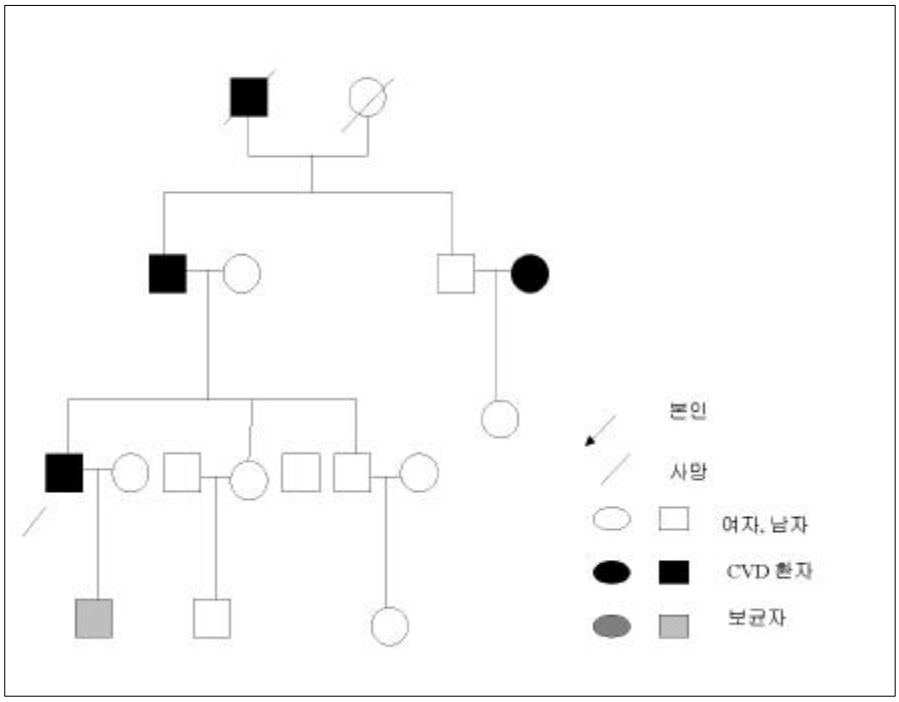
4. 2. 1. 1 조사대상

자료는 S 병원에 내원하여 심혈관질환 판정을 받은 2명을 대상으로 그들 가족의 APOE 유전자의 유전형에 따른 심혈관 질환 발병여부에 대한 가계도

를 조사하였다. 2개의 가계도의 총인원은 44명으로 첫번째 가계도는 27명으로 두번째 가계도는 17명으로 구성되어있다. 이외에 각 개인들의 성별, 나이, 흡연과 음주 여부, 심혈관질환의 위험요인으로 알려진 고혈압, 고지혈증, 당뇨병의 유무 등이 조사되었다. 보균자는 APOE 대립유전자중 질병유전자로 알려져 있는 $\varepsilon 4$ 또는 $\varepsilon 2$를 보유한 사람이다. 두 개의 가계도는 아래 [그림 4-1] [그림4-2]와 같이 구성되어있다.



[그림 4-1] 가계도1



[그림 4-2] 가계도2

4. 2. 1. 2 조사목적 및 방법

실제 자료분석에 사용될 가계도는 표식유전자의 유전자형을 조사하기 전에 질병의 표현형 즉, 질병을 앓고 있느냐의 여부등이 자료이다. 질병유전자 APOE의 대립형질의 모집단빈도와 각 대립형질의 조합에서의 발병률, 표식유전자의 대립형질에 관한 정보 등 조사된 정보를 가지고 각 개인의 질병유전자와 표식유전자의 유전자형을 모의시험 하였다.

이 모의실험된 질병유전자와 표식유전자를 가지고 이 유전자들의 연관성 정도의 검정력을 하는데 조사된 가계도 자료가 연관성 분석을 하기에 충분한 정보를 가지고 있는가를 판단하고 어떠한 표식유전자가 연관성 분석 시 좀더

효율적으로 사용될 수 있는지를 추정해 본다.

자료의 기초 통계량들은 SAS Package를 사용하여 분석하였고 가계도 검정력에 관한 추정은 SIMLINK를 사용하여 모의실험 하였다.

4. 2. 2 임상적 특성

대상 가계도 총 인원 44 명중 12 명(27.3 %)이 심혈관 질환으로 진단되었고 이중 남자는 22명중 7명 (31.8 %), 여자는 22명중 5명(22.7 %) 이었다. 첫번째 가계도 에서는 27명중 8명(29.6%)이 심혈관 질환으로 진단되었고 두번째 가계도 에서는 17 명중 4명(23.5 %)이 진단되었다[표 4-2].

[표 4-2] 심혈관질환 발생빈도

구분 \ 심혈관질환		유		무(보균자 포함)	
		명수	%	명수	%
가계도	1	8	29.6	19	70.4
	2	4	23.5	13	76.5
성별	남	7	31.8	15	68.2
	녀	5	22.7	17	77.3

전체 조사된 사람들의 나이는 7세 부터 80세 까지 분포되어 있고 심혈관 질환이 발병된 사람들의 나이는 38 세부터 80 세까지 분포되어있다[표 4-3].

[표 4-3] 나이의 분포

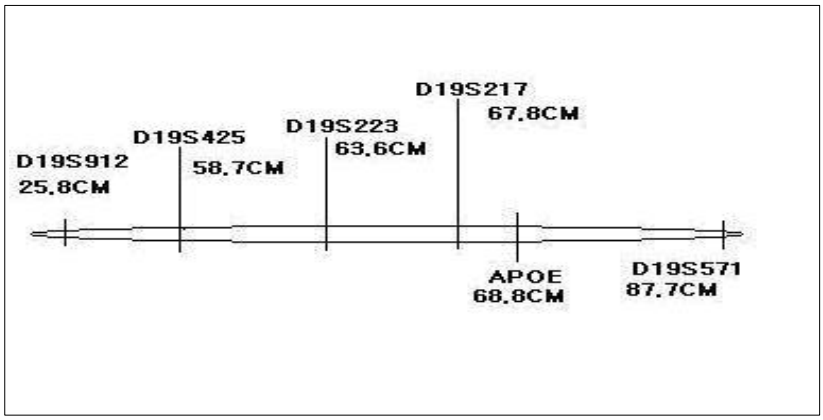
나이 / 심혈관질환	명수	평균	표준편차	최대값	최소값
유	12	56	15.67	38	80
무	32	35.34	19.83	7	78

4. 3 가계도 자료의 정보충분성 측정 결과

4. 3. 1 단일 표식유전자를 사용한 가계도 자료의 경우

4. 3. 1. 1 표식유전자의 대립형질 빈도가 동일한 경우

19번 염색체(Genebank)위에 있는 유전적으로 떨어진 거리가 다른 표식유전자 D19S912 ,D19S425 ,D19S223 ,D19S217 ,D19S571 5개를 이용하여 심혈관질환을 일으키는 APOE 유전자와의 연관성 여부에 대한 가계도의 검정력을 모의실험으로 추정하였다. 모의 실험의 반복횟수는 500번으로 하였고 각 표식유전자의 대립형질은 동일한 빈도를 같는 공우성형질(codominant)을 가정하였다. 각 표식유전자와 APOE 와의 거리는 아래 [그림 4-3]과 같다.



[그림 4-3] APOE 와 표식유전자의 위치1

모의실험을한 결과 표식유전자와 APOE유전자와의 연관성에 대한 $E(Z_{max})$ 와 $P(Z_{max} \geq c)$는 아래 [표4-4] [표4-5] 와 같다.

[표 4- 4] 두 가계도의 $E(Z_{max})$ I

Ture recombination fraction	Pedigree		
	1+2*	1	2
0.00	1.448	0.942	0.532
0.01	1.360	0.027	0.022
0.05	1.092	0.734	0.426
0.1	0.872	0.600	0.369
0.2	0.493	0.334	0.263
0.3	0.312	0.247	0.180
0.4	0.164	0.152	0.117

*Maximum summed lod score for pedigrees 1 and 2 together.

[표 4-5] 두 가계도의 $P(Z_{max} \geq c)$ I

True recombination fraction	Constant	Pedigree		
		1+2*	1	2
0.000	0.5	0.848	0.692	0.450
	1.0	0.680	0.410	0.258
	1.5	0.446	0.206	0.014
	2.0	0.270	0.062	0.000
	2.5	0.122	0.010	0.000
	3.0	0.034	0.000	0.000
0.010	0.5	0.872	0.708	0.396
	1.0	0.628	0.400	0.224
	1.5	0.396	0.162	0.010
	2.0	0.212	0.050	0.000
	2.5	0.076	0.010	0.000
	3.0	0.024	0.000	0.000
0.050	0.5	0.744	0.586	0.356
	1.0	0.494	0.312	0.182
	1.5	0.270	0.124	0.012
	2.0	0.138	0.048	0.000
	2.5	0.060	0.000	0.000
	3.0	0.014	0.000	0.000
0.100	0.5	0.622	0.460	0.294
	1.0	0.366	0.218	0.140
	1.5	0.202	0.104	0.008
	2.0	0.080	0.028	0.000
	2.5	0.028	0.002	0.000
	3.0	0.004	0.000	0.000
0.200	0.5	0.386	0.276	0.192
	1.0	0.160	0.082	0.086
	1.5	0.066	0.018	0.004
	2.0	0.028	0.008	0.000
	2.5	0.008	0.002	0.000
	3.0	0.002	0.000	0.000

*Maximum summed lod score for pedigrees 1 and 2 together. (계속)

[표 4-5] 두 가계도의 $P(Z_{max} \geq c)$ Ⅰ (계속)

True recombination fraction	Constant	Pedigree		
		1+2*	1	2
0.300	0.5	0.234	0.180	0.116
	1.0	0.084	0.058	0.046
	1.5	0.032	0.014	0.006
	2.0	0.004	0.000	0.000
	2.5	0.000	0.000	0.000
	3.0	0.000	0.000	0.000
0.400	0.5	0.110	0.102	0.064
	1.0	0.036	0.034	0.020
	1.5	0.010	0.008	0.000
	2.0	0.000	0.000	0.000
	2.5	0.000	0.000	0.000
	3.0	0.000	0.000	0.000

*Maximum summed lod score for pedigrees 1 and 2 together.

위 [표4-4] 와 [표4-5]를 보면 전체적으로 모든 재조합률에서 가계도 2보다는 가계도 1이 질병유전자와 표식유전자 사이의 연관성 분석에 대한 더 높은 검정력을 보여 주고 있다.

또, APOE와 D19S912 유전자는 유전적 거리가 약 43 cM (centimorgan) 간 떨어져 있는데, 일반적으로 거리가 1cM 떨어져 있을 때 1% 의 유전자 재조합이 일어난다고 알려져 있다. 그러므로 두 유전자사이의 재조합률을 0.4로 볼 수 있다. 이러한 방법으로 각 표식유전자와 AOPE의 재조합률을 고려하여 보면 [표 4-6]과 같다.

[표 4-6] APOE 유전자와 각 표식유전자의 거리에 따른 검정력

표식유전자	APOE 와의 거리 (cM)	$E(Z_{max})$	$P(Z_{max} \geq c)$*
D19S217	1	1.360	0.024
D19S223	5.2	1.092	0.014
D19S425	10.1	0.872	0.004
D19S571	18.9	0.493	0.002
D19S912	43	0.164	0.000

* 특정상수 c 는 3 값으로 비교. 가계도 1과 가계도 2의 합을 사용

위 결과에서 보면 질병유전자 APOE 와의 유전적 거리가 멀수록 검정력은 감소하는 것을 볼 수 있는데 즉, 사용하는 표식유전자와 질병유전자의 거리가 많이 떨어져 있을수록 두 유전자 사이의 연관성을 검정해내는 검정력이 떨어진다고 볼 수 있다. 또, $P(Z_{max} \geq c)$ 와 $E(Z_{max})$ 의 값이 전체적으로 낮은 것으로 볼 때 비록 가계도 1이 가계도 2보다는 더 좋은 검정력을 보이지만([표4-4] [표4-5]) 연관성 분석을 할 경우 가계도의 수 같은 가계도가 가지고 있는 정보의 양이 분석을 하기에는 충분하지 않다고 볼 수 있다.

4. 3. 1. 2 표식유전자의 대립형질 빈도가 다른 경우

19번 염색체위에 있는 유적적 거리가 동일하고 각 대립형질의 빈도가 틀린 표식유전자 snp461과 snp459 이용하여 심혈관질환을 일으키는 APOE 유

전자와의 연관성 여부에 대한 가계도의 검정력을 모의실험으로 추정하였다.

APOE와의 유전적 거리는 두 유전자 모두 1.4cM 떨어져 있고 snp461 유전자는 두 가지의 대립형질로 구성되어 있고 각 대립형질의 빈도는 0.091과 0.909로 구성된다. snp459 유전자도 snp461 와 동일하게 두 가지의 대립형질로 구성되어있고 대립형질의 빈도는 각각 0.49, 0.51이다 (Eden R. Martin et al, 2000).

APOE 와의 유전적 거리가 두 유전자 모두 1.4cM 으로 동일하므로 재조합률 0.01하에서 두 유전자의 $E(Z_{max})$와 $P(Z_{max} \geq c)$을 살펴보면 [표 4-7] [표4-8] 과 같다.

[표 4-7] 두 가계도의 $E(Z_{max})$ Ⅱ

표식유전자	대립형질의 빈도	Pedigree		
		1+2*	1	2
snp459	0.49	1.376	0.910	0.489
snp461	0.091	0.598	0.410	0.193

*Maximum summed lod score for pedigrees 1 and 2 together.

[표 4-8] 두 가계도의 $P(Z_{max} \geq c)$ Ⅱ

표식유전자	대립형질의 빈도	Constant	Pedigree		
			1+2*	1	2
snp459	0.49	0.5	0.876	0.710	0.398
		1.0	0.624	0.412	0.224
		1.5	0.404	0.166	0.010
		2.0	0.226	0.054	0.000
		2.5	0.082	0.010	0.000
		3.0	0.024	0.000	0.000
snp461	0.091	0.5	0.454	0.364	0.136
		1.0	0.272	0.152	0.110
		1.5	0.106	0.042	0.020
		2.0	0.044	0.010	0.000
		2.5	0.044	0.002	0.000
		3.0	0.000	0.000	0.000

*Maximum summed lod score for pedigrees 1 and 2 together.

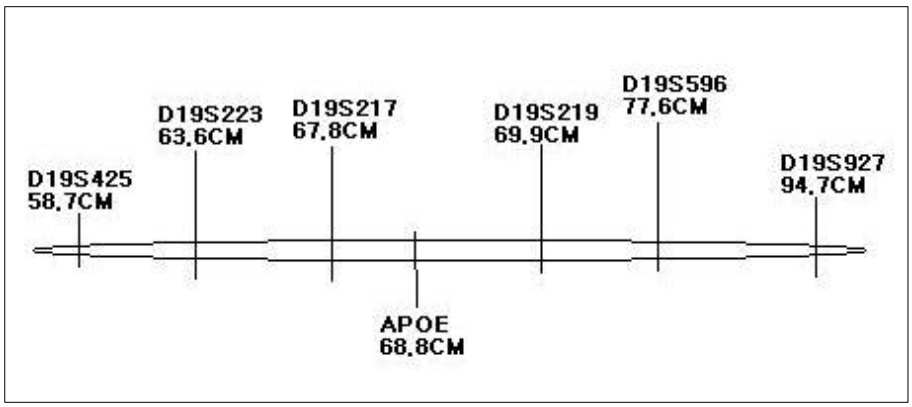
위 [표4-7] 과 [표 4-8]의 결과에서 보듯이 가계도 1이 가계도 2보다 훨씬 높은 검정력을 보이는데 이는 대립형질의 모집단 빈도가 동일한 앞 절의 결과와 일치한다.

또 두 가계도 모두에서 snp459 유전자가 snp461 유전자 보다 약 두 배 정도 높은 검정력을 보이고 있다. 즉, 대립형질의 빈도가 비슷한 표식유전자를 사용할 수 록 가계도 자료의 검정력이 높아진다고 판단 할 수 있다.

4. 3. 2 두개의 표식유전자를 사용한 가계도 자료의 경우

19번 염색체 위에 있는 D19S217, D19S219, D19S223, D19S596, D19S425,

D19S927 (Genebank) 6 개의 표식유전자를 사용하여 질병유전자 APOE 와의 연관성에 대한 검정력을 추정해 보았다. 이 절에서는 APOE를 사이에 두고 있는 두 표식유전자를 동시에 고려하여 검정력을 추정하였다. 각 표식유전자와 APOE 와의 거리는 아래 [그림 4-4] 와 같다.



[그림 4-4] APOE 와 표식유전자의 위치2

표식유전자를 3쌍으로 나누어 [(D19S217, D19S219), (D19S223, D19S596), (D19S425, D19S927)] 연관성 분석에 관한 검정력을 추정한 결과가 [표 4-9] [표 4-10]과 같다.

[표 4-9] 표식유전자의 거리에 따른 $E(Z_{max})$

표식유전자		거리 (cM)	Pedigree		
			1+2*	1	2
D19S217	D19S219	2.1	2.513	1.594	0.927
D19S223	D19S596	29	1.835	1.208	0.722
D19S425	D19S927	36	1.127	0.777	0.526

*Maximum summed location score for pedigrees 1 and 2 together.

[표 4-10] 표식유전자의 거리에 따른 $P(Z_{max} \geq c)$

표식유전자	거리 (cM)	Constant	Pedigree		
			1+2*	1	2
D19S217 D19S219	2.1	0.5	0.97	0.914	0.682
		1.0	0.916	0.772	0.588
		1.5	0.838	0.566	0.174
		2.0	0.712	0.366	0
		2.5	0.526	0.128	0
		3.0	0.324	0.004	0
D19S223 D19S596	29	0.5	0.898	0.792	0.56
		1.0	0.782	0.606	0.412
		1.5	0.626	0.362	0.08
		2.0	0.446	0.172	0
		2.5	0.276	0.036	0
		3.0	0.132	0.002	0
D19S425 D19S927	36	0.5	0.718	0.618	0.43
		1.0	0.526	0.338	0.252
		1.5	0.336	0.148	0.02
		2.0	0.164	0.042	0
		2.5	0.074	0	0
		3.0	0.02	0	0

*Maximum summed location score for pedigrees 1 and 2 together.

위 [표4-9] 과 [표 4-10]의 결과에서 보듯이 가계도 2 가 가지고 있는 자료의 정보 보다는 가계도 1 의 정보가 두 유전자 사이의 연관성에 더 큰 검정력을 보인다. 또, 두 가계도 모두에서 표식유전자 사이의 거리가 근접 할

수록 질병유전자와 표식유전자 사이의 연관성에 대한 검정력이 높았다.

이 결과는 앞 절의 단일 유전자를 사용한 검정의 결과와 같다. 그러나 동일한 가계도 자료로 연관성분석을 할 경우 두 유전자를 사용하여 검정하는 것이 단일 유전자를 사용하여 검정하는 것 보다 좀더 높은 검정력을 보인다.

제 4장 결론 및 고찰

지금까지 통계적 검정력을 통한 가계도 자료의 정보 충분성 추정에 관한 이론적인 방법들을 고찰 해보았고 이와 관련하여 연관성 분석의 의미와 연관성 분석에서의 불완전 침투의에 관하여 알아보았다.

가계도 자료의 정보 충분성 추정은 가계도 자료 수집 후 표식 유전자의 유전형 판독 전에 시행하는 분석으로서 필요한 가계도의 수나 발병된 구성원과 같은 분석시 필요한 가계도 자료와 적절한 표식유전자를 제시하여 줌으로서 실제 연관성 분석에서 표식유전자 판독에 소요되는 막대한 시간과 비용을 좀더 효율적으로 배분할 수 있게 해주고 찾고자 하는 질병 특성유전자의 연관성정도의 검정력을 높일 수 있다.

일반적으로 가계도 자료의 정보 충분성 측정은 Lod Score를 이용하는 방법과 Fisher's Expected information을 이용하는 방법으로 나누어 볼 수 있는데 본 논문의 실제 심혈관 자료를 이용한 정보 충분성 측정에서는 Lod Score를 이용한 $P(Z_{max} \geq c)$ 와 $E(Z_{max})$를 사용하였다.

또한, 정보충분성 검정전에 관련된 이론을 살펴보았는데 우선 연관성 분석이란 두 유전자 즉, 표식유전자와 질병유전자의 연관정도를 측정하는 것인데 분석의 가장 큰 목적은 이미 알려진 표식유전자를 사용하여 질병과 관련된 특성유전자를 찾아내는데 있다. 대부분의 질병에서 연관성분석을 할 경우 표현형과 유전형이 일치하지 않는 불완전 침투가 많이 발생는데 이 불완전 침투에서의 특성 유전자를 찾을 때에는 재조합률을 추정하는 것 외에 질병의 침투율도 추정해야 한다. 이 두 가지 모수를 추정하는 방법에는 크게 두가지 방법이 있는데 우도함수를 이용한 추정과 질병의 위험률을 이용한 추정이다.

우선 우도 함수를 이용한 추정은 모든 자손들을 대상으로 재조합률과 침

투율을 동시에 추정해야 하는데, 이 두 모수를 동시에 추정하는 것은 현실적으로 그리 쉬운 문제가 아니다. 또 그 추정의 대상을 모든 자손을 대상으로 하기 때문에 현실적 비효율적인 부분이 있다. 이러한 문제점들 때문에 질병의 발병률은 통한 추정 방법이 대두되었는데 이 방법은 결합우도를 사용하여 침투율을 추정하는 대신에 발병된 사람과 발병되지 않은 모든 사람을 대상으로 질병의 발병률을 추정하고 이를 이용하여 침투율의 추정한 다음 발병된 사람들만을 대상으로 재조합률을 추정한다. 우도와 발병률의 장점을 이용한 추정으로 앞의 여러 가지 문제점들을 해결할 수 있었다.

실제 자료를 이용한 자료의 가계도 자료의 검정력 추정에서는 일 병원에 내원한 심혈관 질환을 가진 두 명의 환자의 심혈관 질환 발생 유무에 따라 조사된 그들의 가계도를 자료를 사용하여 자료의 충분성에 관한 검정력을 추정해 보았다.

동일한 가계도 에서 단일 표식 유전자를 사용한 추정과 두개의 표식유전자를 동시에 고려한 추정을 해보았는데, 두 추정 모두에서 27 명중 8명이 심혈관 질환으로 판명된 첫번째 가계도(그림 3-1)가 가지고 있는 정보의 양이 두번째 가계도(그림 3-2)보다 연관성을 판단할 수 있는 검정력이 높았다. 또, 두 유전자 사이의 거리가 가까울 수록 즉, 단일 유전자의경우 질병유전자와 표식유전자와의 거리가 가까울 수록 두 표식 유전자의 경우 두 표식유전자의 거리가 가까울 수록 가계도 자료에 대한 검정력이 높았다.

동일한 가계도일 경우 단일 표식유전자를 사용할 때 보다는 두 개의 표식유전자를 사용할 때가 검정력이 더 좋았고 단일 표식 유전자에서 표식유전자의 각 대립형질의 빈도가 비슷할 수록 검정력이 높게 나왔다.

그러나 대체적으로 모든 경우 가계도 자료의 정보 충분성에 관한 검정력이 낮게 도출되었는데, 이는 실제 이 자료로 연관성 분석을 할 경우 가계도의 수 나 각 가계의 구성원의 수 같은 가계도 자료가 좀 더 필요하다고 판단된다. 그리고 분석에 필요한 적절한 표식유전자 선택 시 표식유전자의 각 대

립형질의 빈도가 비슷하고 질병유전자와의 유전적 위치가 근접한 유전자 일수 록 검정의 효율성을 높일 수 있을 것이라 생각된다.

이와 더불어 가계도의 검정력이 전체적으로 낮게 나온 이유는 심혈관 질환이 여러 유전자가 복합적으로 작용하여 발병되는 복잡한 유전적 경향을 띄는 질환으로서 APOE 질병유전자 이외에 다른 질병유전자를 고려할 필요성이 있다고 판단된다.

참고문헌

강동민역. 래리고닉. 마트휠리스, 생물 유전학 길잡이, 도서출판국제, 1995

박상규. 남정원. 이은정. 로버트 폴락, 생명의 기호, 민음사, 1996

박훈기, 한국인 중년 남성에서 심혈관질환의 발생 위험요인에 관한 전향적 코호트 연구 , 1999, 서울대학교 대학원 의학과

이정우 역. 프랑수아 쟈콥, 생명의 논리 유전의 역사, 민음사, 1994

이철훈 역. Mark Wheelis, 유전학으로의 산책, 도서출판 정담, 1993

정재언 역. Steve Jones, 유전학 무엇이 세계를 움직이는가?, 이두아이콘 총서, 1995

정해영 역. 이쿠타사토시, 유전자와 DNA 의 최첨단, 월드사이언스, 2000

통계청, 사망원인통계연보, 1999

한국보건사회연구원, 국민건강 영양조사, 1998

한국보건사회연구원. 보건복지부, 국민건강 영양조사-건강면접조사 부문, 1998

한국유전학회 역. 죠셉 레빈. 데이비드 스키, 유전자 생명의 원천, 전파과학사, 1996

Boehnke, M., Estimating the power of a proposed linkage study., *Am. J. Hum* , 1986 : 39 ; 513 - 27

Boehnke M., Ploughman, L.M., Estimating the power of a proposed linkage study for a complex genetic trait., *Am. J. Hum.*, 1989 : 44 ; 543 - 51

Carlos Lahoz. et al., Apolipoprotein E genotype and cadiovacular disease in the framingham heart study., *Atherosclerosis*., 2001 : 154 ; 529 - 37

De Vries RRP et al, HLA-linked genetic control of host response to mycobacterium leprae., *The Lancet*., 1976 : 18 ; 1328.

Dirk F. Moore, Nilanjan Chatterjee, David Pee and Mitchell H. Gail., Pseudo-Likelihood Estimates of the cumulative Risk of an Autosomal Dominant Disease Form a Kin-Cohort Study., *Genetic Epidemiology*, 2001 : 20 ; 210-27

Douglas F. Easton, Deborah Ford, D. Timothy bishop., Breast and Ovarian Cancer Incidence in BRCA1-Mutation., *Am. J. Hum.*, 1995 : 56 ; 265-71

Eden R. Martin. et al., SNPing away at complex disease : analysis of single-nucleotide polymorphisms around apoe in alzheimer disease., *Am. J. Hum.*, 2000 : 67 ; 383-94

Elandt-Johnson, R.C., Probability Models and Statistical Methods in Genetics., *New York John Wiley*, 1971

Elston, R. C., Ascertainment and age of oneset in pedigree analysis., *Hum. Hered.*, 1973 : 23 ; 105 - 12

Fisher, R. A., Theory of statistical estimation., *Proc. Comb. Philos. Soc.*, 1925 : 22 ; 700-725

Green JR, Low HC, Woodrow JC, Inference of inheritance of disease using reptition of HLA haplotypes in affect ed siblings., *Ann Hum Genet.*, 1983 : 47; 73-82

Haldans JBS., The combination of linkage values and the calculation of distances between the loci of linked factors., *Journal of Genetics.*, 1919 : 8 ; 299-309

Haseman, J. K. and R. C. Elaton., The investigation of linkage between a quantitative trait and a marker locus., *Behav. Genet.* , 1972 : 2 ; 3-19

Jurg ott, Analysis of Human Genetic Linkage, *The Johns Hopkins University Press Baltimore and London* , 1991

Lander, E. S. and D. Botstein, Strategies for studying heterogeneous genetic trait in humans by using a linkage map of restriction fragment length polymorphims., *Proc. Natl. Acad. Sci. USA* ., 1986a : 83 ; 7353-57

Lange K. and Boehnke M. Extension to pedigree analysis. V. optimal calculation of mendelian likelihoods., *Hum. Hered.*, 1983 : 33 ; 291-301

Lange, K. , The affected sib-pair method using identity by state relations., *Am. J. Hum.*, 1986 : 39 ; 148-50

Lange K. and Boehnke M., Weeks D. Programs for pedigree analysis : MENDEL, FISHER and dGENE., *Genetic Epidemiology*, 1988 : 5 ; 471-72

Mitchell H. Gail, David Pee, Jacques Benichou and Raymond Carroll, Designing Studies to estimate the Penetrance of an identified autosomal Dominant mutation., *Genetic Epidemiology*, 1999 : 16 ; 15-39

Mitchell H. Gail, David Pee, Raymond carroll., Kin-Cohort Designs for Gene Characterization., *Journal of the National Cancer Institute Monographs*., 1999: 26.

Morton NE. Sequential tests for the detection of linkage., *Am. J. Hum.*, 1955 : 7 ; 277 - 318.

Penrose LS.,The detection of automal linkage analysis., *Ann Eugenics*., 1935 : 6 ; 133-8

Risch, N. Linkage detection tests under heterogeneity., *Genetic Epidemiology*, 1989 : 6; 473-80.

Risch, N. Linkage strategies for genetically complex traits. **I**. Multilocus Models. *Am. J. Hum.*, 1990 : 46 ; 222 - 28

Risch, N. Linkage strategies for genetically complex traits. **II**. Multilocus Models. *Am. J. Hum.*, 1990 : 46 ; 229 - 41

Risch, N. Linkage strategies for genetically complex traits. **III**. Multilocus

Models. *Am. J. Hum.*, 1990 : 46 ; 242 - 53

Rossen, R. D., E.J. Brewer, R.M. sharp., J. Ott, and J.W. Templeton., Familial rheumatoid arthritis: Linkage of HLA to disease susceptibility locus in four families where proband presented with juvenile rheumatoid arthritis., *J. Clin .Invest.* 1980 : 65 ; 629 - 42

Smith CAB. Testing for heterogeneity of recombination fraction values in human genetics., *Annals of Human Genetics*, 1963 : 27 ; 175 - 82

Thompson, E. A., C. Cannings, and M .H. Skolnic, Ancestral inference. I. The problem and the method., *Ann. Hum. Genet.* : 1978a :42 ; 95-108.

http://www.ncbi.nml.nih.gov/genemap99

부록 1. 유전용어의 표기법

• allele	대립유전자 또는 대립형질
• autosome	상염색체
• carrier	보균자
• chromosome	염색체
• codominant	공우성형질
• codon	코돈
• diploid	이배체
• domonant	우성형질
• exon	엑숀
• gene	유전자
• gentic marker or marker gene	표식유전자
• genome	게놈
• genotype	유전형
• haploid	반수체
• Hardy-weinberg Equilibrium	HWE
• intron	인트론
• locus	로쿠스
• map distance	유전적 거리
• map function	사상함수
• meiosis	감수 분열
• monozygotic twin (MZ twin)	일란성 쌍둥이
• mutation	돌연변이
• nucleotide	뉴클레오타이드
• penetrance	침투율

• phenotype	**표현형**
• polymorphism	**다형성**
• recessive	**열성형질**
• recombination	**재조합**
• recombination fraction	**재조합률**
• trait	**형질**
• transcription	**전사**
• X- chromosome	**성염색체**

부록 2. 유전용어의 해설

• 감수분열(또는 생식세포분열:meiosis) : 유성생식을 하는 생물이 생식세포를 형성할 때 일어나는 핵분열. 즉 염색체의 수가 절반인 딸세포를 만드는 과정.

•게놈(genome) : DNA 로 표현되는 한 세포 또는 바이러스 전체유전정보.

•공우성형질(codominant) : 두 개의 대립형질이 서로 그 특성을 나타내는 것.

•뉴클레오타이드(nucleotide) : DAN 또는 RNA 사슬의 기본적인 단위. 이들은 염기(아데닌, 구아닌, 시토신, RNA에선 우라실로 대체되는 티민으로 구성되어 있다)와 당으로 구성되어 있으며 인산에 의해 연결된다.

•다형성(polymorphism) : 다른 로쿠스 위에 두 개이상의 대립형질이 나타나는 경우. 이 중에서 로쿠스에 나타나는 대립형질이 모두 다른 경우를 완전다형성(fully polymorphism) 이라 한다.

•대립형질(또는 대립유전자: allele) : 염색체의 같은 위치에 존재하는 유전자인데 두 가지 이상의 형태(alternative forms)가 존재하는 것.

•돌연변이(mutation) : 유전물질의 구조에 일어난 변화(즉, DNA의 염기서열에 일어난)로 유전된다. 돌연변이는 유전자의 기능을 변화시킨다.

•DNA(deoxyribon-ucleic acid) : 디옥시리보핵산의 약칭. 주로 세포 핵 속에 있는 중요한 유전물질로서 2중 나선의 고분자 구조를 가지고 2중 나선 이합에 의하여 복제가 가능하다.

•Lod Score : 연관성 분석에서 연관성 가설을 검정할때 사용하는 검정통계량으로서, 우도비에 로그값을 취한 값이다. 일반적으로 상염색체의 경우 최대 Lod Score 가 3보다 크면 연관 되어있다고 판단한다.

•로쿠스(locus, 복수형 loci) : 유전자의 염색체 상에서의 위치.

•보균자(carrier) : 질병을 유발하는 유전자를 보유한 개체.

•반수체(haploid) : 체세포가 이배체 상태를 유지하기 위하여 수정할 때 결합하는 생식세포나 배우자의 핵에는 한개의 염색체만이 존재하는데 이러한 핵을 가진 세포 또는 개체를 반수체 라고 한다.

•사상함수(map function) : 추정된 재조합률을 실제 유적적 거리(map distance)로 나타내주는 함수.

•상염색체(autosome) : 성염색체를 제외한 염색체를 말한다.

•성염색체(X-chromosome) : 성을 결정짓는 염색체로 남성은 X 염색체를 1개 가지고 있고 여성은 두 개를 가지고 있다.

•Hardy-Weinberg Equilibrium(HWE) : 대립형질에 따른 유전형의 모집단 비율이 세대가 바뀌어도 일정하게 유지된다는 성질.

•엑쇤(exon) : DNA에서 유전적 기능을 담당하는 부분으로 다백질을 지어하는 유전자의 한 부분을 구성한다. 정보의 지정능력이 없는 인트론 사이사이에 흩어져 있다.

•열성형질(recessive) : 두 개의 대립형질 중에서 유전자 특성이 표현되어지지 않는 것 .

•염색체(chromosome) : 세포의 핵분열 때 핵 내의 유전자가 포함되어 있는 염색질이 모여서 이루어진 작은 막대기 모양의 소체.

•우성형질(dominant) : 두 개의 대립형질 중에서 유전자 특성이 표현되어 지는 것.

•유전자(gene) : 하나의 기능을 가지는 산물을 지정하는 DNA 서열로서 유전형질을 규정하는 인자이다. 유전자들은 유전현상의 기본단위이다.

•유전형(genotype) : 생물체 개체의 특성을 결정짓는 유전자의 결합양식.

•유전적거리(map distance) : 염색체 상에 위치한 유전자들의 거리로서 측정단위로는 주로 cM(centi Morgan)을 사용하며 보통 1cM 떨어져 있을때 1%의 재조합이 일어난다고 한다.

•이배체(diploid) : 하나의 세포에 적용하였을 때 두 개씩의 완전한 염색체 세트를 갖는 세포.

•Expected Lod Score(ELOD) : Lod Score 의 가중평균.

•인트론(intron) : 진핵세포에서 유전자를 구성하는 엑쇤의 사이사이에 놓이는 DNA로서(junk DNA) 정보를 지정하지 않은 염기들로 현재까지 알려진 기능이 없다. mRNA를 합성하는 동안 인트론도 함께 RNA로 전사

(transcription)가 되지만 잇기(splicing)과정에서 없어진다.

•일란성쌍둥이(monozygotic twin: MZ twin) : 부모로부터 받은 유전적인 요소가 동일한 사람.

•재조합(recombination) : 세포 분열과정 중 염색체의 교차(cross over)로 인한 염색체 분절(segment)의 교환으로 새로운 조합의 대립형질(allele)을 가지는 것.

•재조합률(recombination fraction) : 재조합이 일어날 확률.

•전사(transcription) : DNA에서 상보적인 RNA를 합성하는 과정으로 유전정보를 이용하는 기본적인 단계.

•침투(침투율: penetrance) : 유전형에 의한 표현형의 조건부 확률로서 유전형과 표현형이 일치할 경우를 완전침투라 하고 일치하지 않을 경우를 불완전침투라 한다.

•코돈(codon) : mRNA에서 하나의 아미노산을 지정하는 세 개의 연속적인 염기.

•표식유전자(genetic marker또는 marker gene) : 염색체에서 위치가 알려져 있는 유전자.

•표현형(phenotype) : 유전자에 있는 유전정보가 그 생물의 형질로서 나타나는 것.

•형질(trait) : 생물의 표면적인 성질, 즉 표현형. 외관뿐만 아니라 체내의 여러 기관의 형태, 기능의 차이 등도 포함된다. 생물의 몸길이, 체중, 여러 기관의 크기처럼 계측되는 형질을 수량적 형질, 형상이 다른 형태적 형질, 생태반응에 관한 것을 생리생화학적 형질이라 한다.

ABSTRACT

The measurement of the informativeness on a set of pedigree data in genetic linkage analysis

Park, Yun Ju

Dept. of Biostatistics and Computing

The Graduate School

Yonsei University

One of the reasons for designing this study is because it is possible to minimize loss of time and cost with the test for sufficiency of information in pedigree by using statistical power since it is examined before genotyping is taken in linkage analysis.

In this thesis we introduce the theory of testing for sufficiency of information on a set of pedigree data, which is done before genotyping of genetic marker is made. And also we describe a method for

estimation of recombination fraction and penetrance.

In order to estimate the power of pedigree we used two pedigrees with cardiovascular disease(CVD) patients obtained a hospital, and examined the number of CVD patients in each pedigree.

Of the two pedigrees, 8 cardiovascular disease patients of 27 family members and 4 of 17 family members were involved in the first and the second family respectively.

We tried to estimate the sufficiency of information by calculating expected maximum LOD score and the probability for maximum LOD score not less than a certain constant c, in the two pedigrees.

In conclusion, we found the first pedigree is more powerful than the second one in analysis with single marker and the estimation method with the flanking marker. In addition to that, we found the fact that the statistical power is increased as the distance in the flanking marker increases. And also, analysis with the flanking marker showed higher statistical power than that with a single marker.

The statistical power is increased as the frequencies of alleles of genetic marker are similar with each other.

Key words : informativeness, power of pedigree, linkage analysis, incomplete penetrance, recombination fraction, penetrance