응용통계연구 제 **21**권 *1*호, *2008*년, *pp. 81–94*

형제 및 자매의 유전자형 자료에 기초한 전달불균형 검정법에 관한 연구*

김진흠[1] 장양수[2]

요 약

전달불균형 검정법(transmission and disequilibrium test)과 같은 가계중심(family-based) 검정법들은 질병 관련 유전자를 찾는 데 매우 유용한 방법으로 알려져 있다. 사례-대조군 연구와 달리 가계중심 검정법들은 집단 혼합(population admixture)으로 인한 영향을 받지 않기 때문에 질병 관련 유전자와 표지자(marker) 사이의 집단 혼합으로 인한 가짜 연관성(spurious association)에 노출될 위험이 없다. 가계중심 검정법들은 대체로 표지자에 대한 부모의 유전자형(genotype) 정보를 필요로 한다. 그러나 고령층에서 발병하는 질병의 경우에는 발단자(proband) 부모의 유전자형을 구할 수 없는 상황에 종종 마주치게 된다. 본 논문에서는 이런 어려움을 극복하기 위해 부모의 유전자형 대신 질병에 노출되지 않은 발단자 형제나 자매의 유전자형을 이용한 검정법을 제안하고자 한다. 이를 위해 먼저 가능한 모든 일배체형(haplotype)에 대해 Mantel-Haenszel 형태의 통계량을 정의하고 그것에 기초한 두 가지 검정통계량을 제안하였다. 모의실험 결과, 제안한 검정법은 집단 혼합으로부터 로버스트하고 유전 양식(mode of inheritance)에 관계없이 상대위험(relative risk)이 증가함에 따라 단조적으로 증가하는 검정력을 갖는 것으로 나타났다. 제안한 검정법을 연세대학교 심혈관계질환 유전체연구센터로부터 수집한 자료에 적용하고 그 결과를 고찰하였다.

주요용어: 순열 검정, 안지오테시노겐 유전자, 연쇄 및 연관성 분석, 일배체형, 전달불균형 검정법.

1. 서론

현재 미국을 비롯한 여러 선진 외국에서는 고혈압, 당뇨 등과 같은 다인성 질환과 관련된 유전자의 위치를 찾는 데 많은 투자와 노력을 기울이고 있다. 그중 하나는 질병 관련 후보 유전자와 질환 사이의 연쇄 및 연관성을 검정하기 위한 통계적 방법론을 개발하는 것이다. 이를 위해 이제까지는 단일 표지자에 의한 연구가 주류를 이루었는데 최근 들어서는 재

* 이 논문은 2006년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행한 연구임(KRF-2006-312-C00087).

1) (445-743) 교신저자. 경기도 화성시 봉담읍 와우리 산 2-2, 수원대학교 자연과학대학 통계정보학과, 부교수. E-mail: jinhkim@suwon.ac.kr

2) (120-743) 서울시 서대문구 신촌동 134번지, 연세대학교 의과대학 심혈관계질환 유전체연구센터, 교수. E-mail: jangys1212@yuhs.ac

조합이 거의 없을 정도로 아주 강하게 연결되어 있는 여러 표지자들의 모임인 일배체형에 의한 연구로 옮겨가고 있다. 이와 같은 연구가 가능하게 된 것은 일차적으로 근거리 내에 많은 표지자들에 대한 유전자형 확정(genotyping)이 가능해졌기 때문이며, 통계적인 측면에서 보았을 때는 표지자들이 서로 강하게 연쇄되어 있기 때문에 단일 표지자에 기초한 연구보다는 일배체형에 기초한 연구가 더 많은 유전정보를 담아 낼 수 있어 통계량의 검정력을 향상시키는 효과가 있기 때문이다.

본 논문에서 다루고자 하는 자료는 발단자의 유전자형과 함께 발단자의 형제나 자매(sib)의 유전자형 정보는 알려져 있지만, 발단자 부모의 유전자형 정보는 결측된 가계자료이다. 이와 같은 자료는 보통 고령층에서 발생하는 질환에서 종종 찾아 볼 수 있는데, 발단자가 고령층에 속해 있다면 그 발단자의 부모는 대체적으로 생존해 있을 가능성이 희박하기 때문에 발단자 부모의 유전자형은 결측될 수 밖에 없다. 본 논문에서는 부모의 유전자형은 결측되었지만 발단자의 형제나 자매의 유전자형 정보를 써서 질병 관련 유전자와 질병의 연쇄 및 연관성 분석을 위한 통계적 방법을 제안하고자 한다.

2절에서는 Spielman과 Ewens (1998)가 제안한 sib-TDT(transmission and disequilibrium test) 방법을 일배체형에 기초한 경우로 확장한 통계량을 제안하고자 한다. 3절에서는 다양한 형태의 모의실험을 통해 제안한 검정통계량의 소표본 성질을 살펴보고자 한다. 4절에서는 제안한 검정법을 연세대학교 심혈관계질환 유전체연구센터로부터 수집한 자료에 적용하여 AGT(angiotensinogen) 유전자와 고혈압의 연쇄 및 연관성 분석을 수행하고 그 결과를 고찰하고자 한다.

2. 결측을 포함하는 자료를 이용한 sib-TDT

2.1. 배경 및 기호

발단자의 연령이 높으면 발단자 부모의 유전자형을 함께 채취하는 것이 종종 불가능하다. 이와 같이 발단자 부모의 유전자형 자료는 결측되었지만 발단자의 형제나 자매의 유전자형을 얻을 수 있다면, 그 sib(s)들이 관심 있는 질병을 갖고 있든지(affected), 갖고 있지 않든지(unaffected)에 관계없이 sib(s)의 추가 정보를 질병 관련 유전자와 질병의 연쇄 및 연관성 분석에 유용하게 활용할 수 있다. 이와 관련된 연구는 Spielman과 Ewens (1998)에 의해 처음으로 시작되었는데, 그들은 본 논문에서 다루고자 하는 일배체형 자료 대신에 유전자 좌위(locus)가 한 개인 단일 표지자에 대해 다루었다.

Spielman과 Ewens (1998) 방법의 기본적인 아이디어는 affected sib(s)의 대립인자(allele) 분포와 unaffected sib(s)의 대립인자 분포를 비교하는 것이다. 즉, 만일 표지자가 질병 유전자와 연관되어 있지 않다면 두 분포가 서로 유사하고, 그렇지 않다면 둘은 서로 상이한 양상을 띨 것으로 기대되기 때문이다. Spielman과 Ewens (1998)의 sib-TDT 방법에서는 부모의 유전자형을 모르기 때문에 sib(s)에 포함된 대립인자의 분포를 서로 비교하는 반면에, Zhao 등 (2000)과 김진흠 등 (2004, 2005)에서 소개한 TDT 방법은 발단자에게 전달되는 부모의 대립인자의 빈도를 비교하기 때문에 두 방법은 서로 다르다고 말할 수 있다. 그러나 두 방법 모두 가계에만 의존하여 통계량을 구축하기 때문에 집단 혼합과 같은 교락요인으

표 2.1: 집단 혼합에 의한 가짜 연관성의 예제: 두 집단이 1:1로 혼합되었을 때

	집단1			집단2			혼합 집단		
표본	M_1	M_2	합계	M_1	M_2	합계	M_1	M_2	합계
사례군	9	1	10	25	25	50	34	26	60
대조군	81	9	90	25	25	50	106	34	140
합계	90	10	100	50	50	100	140	60	200
	$\chi^2 = 0(p$값 $= 1.000)$			$\chi^2 = 0(p$값 $= 1.000)$			$\chi^2 = 7.26(p$값 $= 0.007)$		

로부터 로버스트한 장점을 지니고 있다. 표 2.1은 가상의 예제로 연관성 연구에서 집단 혼합이 어떻게 교락요인으로 작용할 수 있는지를 보여주고 있다. 집단 혼합 이전에는 대립인자 M_1과 M_2의 분포가 각각 '집단1'에서는 '9:1', '집단2'에서는 '1:1'로 사례군과 대조군에서 모두 동일하고, 검정통계량의 유의확률이 모두 1이기 때문에 후보 유전자와 질병 간에 아무런 연관성이 없다. 그럼에도 불구하고 두 집단이 '1:1'로 혼합된 이후에는 사례군과 대조군의 대립인자 분포가 같지 않을 뿐만 아니라 검정통계량의 유의확률도 0.007으로 매우 작기 때문에 후보 유전자와 질병 간에 연관성이 있는 것으로 결론짓는 오류를 범할 수 있다. 일반적으로 사례-대조 연구와 같은 집단중심(population-based) 검정법들은 이와 같은 교락요인으로부터 자유롭지 못한 반면에, TDT 연구와 같은 가계중심 검정법들은 가계 내에서 표지자들의 전달/비전달(transmission/non-transmission)을 바탕으로 하기 때문에 표지자와 질병 유전자가 실재로는 관련이 없음에도 불구하고 다른 원인에 의해 둘이 서로 연관되어 있는 것으로 결론 내리는 오류를 피할 수 있는 장점을 지니고 있다.

본 논문에서는 단일 표지자에 대해 다룬 Spielman과 Ewens (1998)과 Monks 등 (1998)의 sib-TDT 통계량을 표지자들의 모임인 일배체형으로 확장하고자 한다. 단일 표지자의 경우와 다르게 이 과정 속에서 가장 먼저 대두되는 어려움은 발단자와 그의 sib(s)에 대한 유전자형의 정보만 알려져 있을 뿐, 일배체형 쌍에 대한 정보는 모른다는 것이다. 이처럼 일배체형 쌍의 불확실성은 유전자 좌위가 많아질수록, 이형접합성 표지자(heterozygous marker)가 많아질수록 지수적으로 증가하기 때문에 본 논문에서 다루고자 하는 구조는 Spielman과 Ewens (1988)과 Monks 등 (1998)에서 다룬 단일 표지자의 경우보다 조금 더 복잡하다고 말할 수 있다. sib-TDT 에서 가장 중심이 되는 통계량은 가계별 affected sib(s)의 일배체형의 개수인데, 만일 모든 sib(s)의 일배체형 쌍을 알 수 있다면 통계량을 계산하는데 어려움이 없을 것이다. 그러나 대부분 sib(s)의 일배체형 쌍이 불확실하기 때문에 일배체형의 개수에 대한 추정이 뒤따라야 할 것이다. 이를 위해 본 연구에서는 Zhao 등 (2000)과 김진흠 등 (2004, 2005)에서 제안한 방법을 도입하고자 한다.

서로 다른 $c(>0)$개의 유전자 좌위에서 이대립인자(bi-allele)를 갖는다고 하자. 이때 총 3^c개의 서로 다른 유전자형이 존재할 수 있으며 이를 $G_1, \ldots, G_k$ 라고 놓자. 단, $k = 3^c$ 이다. 서로 다른 F개의 가계가 있다고 하자. $f(f = 1, \ldots, F)$번째 가계의 sib(s)는 총 N_f명이며, 그중에서 affected sib(s)는 N_f^a명 이고 unaffected sib(s)는 N_f^u명이라고 하자. 또한 f번째 가계의 sib(s)들 중에서 유전자형이 $G_g(g = 1, \ldots, k)$인 sib(s)는 총 t_{fg}명 이며, 그

표 2.2: f번째 가계의 sib(s)들의 질병의 상태와 유전자형에 따른 분할표

	유전자형				
질병 상태	G_1	G_2	...	G_k	합계
affected	x_{f1}	x_{f2}	...	x_{fk}	N_f^a
unaffected	y_{f1}	y_{f2}	...	y_{fk}	N_f^u
합계	t_{f1}	t_{f2}	...	t_{fk}	N_f

중에서 affected sib(s)는 x_{fg}명, unaffected sib(s)는 y_{fg}명이라고 하자. 표 2.2는 f번째 가계에서 질병의 상태(affected/unaffected)와 유전자형에 따른 sib(s)의 수에 대한 분할표이다. $N_f^a, N_f^u, \mathbf{t}'_f = (t_{f1}, \ldots, t_{fk})$ 가 주어졌을 때 표지자와 질병이 연쇄되어 있지 않다는 귀무가설(H_0) 하에서 표 2.2은 다변량 초기하분포를 따른다. 한편 표 2.2의 자료구조에서 Spielman과 Ewens (1998)가 언급한 것처럼 다음 두 조건이 만족되어야 한다.

- 조건1: 가계별로 적어도 한 명의 affected sib(s)와 unaffected sib(s)가 존재해야 한다. 즉, $N_f^a \geq 1, N_f^u \geq 1$이다.
- 조건2: 가계별로 모든 sib(s)의 유전자형이 같지 않아야 한다.

2.2. 일배체형 쌍을 안다고 가정할 경우

먼저 발단자를 포함하여 모든 sib(s)의 일배체형 쌍에 대한 정보를 알고 있다고 가정하자. 이대립인자를 갖는 c 개의 유전자 좌위에서 가능한 서로 다른 일배체형들을 $h_1, \ldots, h_l$ 라고 놓자. 단, $l = 2^c$ 이다. f번째 가계에 속한 N_f 명의 sib(s)들 중에서 $hh(h = h_1, \ldots, h_l)$ 일배체형 쌍을 갖는 sib(s)의 수와 $hm(m \neq h)$ 일배체형 쌍을 갖는 sib(s)의 수를 각각 r_{fh}, s_{fh} 라 하자. 모든 sib(s)의 유전자형에 대한 일배체형 쌍을 알고 있기 때문에 r_{fh}, s_{fh}는 유일하게 결정된다. f번째 가계의 affected sib(s)들로부터 관측된 일배체형 h의 개수를 O_{fh}로 정의하자. 일배체형 h에 대해 상수벡터 $\mathbf{c}'_h = (c_{h1}, \ldots, c_{hk})$ 라고 하자. 여기서 c_{hg}는 유전자형 G_g에 포함된 일배체형 h의 개수에 따라 2, 1, 혹은 0의 값을 갖는다. $\mathbf{x}'_f = (x_{f1}, \ldots, x_{fk})$ 라고 하면 $O_{fh} = \mathbf{c}'_h \mathbf{x}_f$와 같이 표현되므로, H_0 하에서 O_{fh}의 평균과 분산은 $\mathbf{x}'_f = (x_{f1}, \ldots, x_{fk})$가 다변량 초기하분포를 따른다는 사실로부터 다음과 같이 유도할 수 있다 (Bishop 등, 1975).

$$E_{fh} = \mathrm{E}(O_{fh}|H_0) = (2r_{fh} + s_{fh})\frac{N_f^a}{N_f},$$

$$V_{fh} = \mathrm{Var}(O_{fh}|H_0) = \{4r_{fh}(N_f - r_{fh} - s_{fh}) + s_{fh}(N_f - s_{fh})\}\frac{N_f^a N_f^u}{N_f^2(N_f - 1)}.$$

일배체형 h에 대해 표준화된 통계량 $z_h (h = h_1 \ldots, h_l)$를 다음과 같이 정의하자.

$$z_h = \frac{\sum_{f=1}^F O_{fh} - \sum_{f=1}^F E_{fh}}{\sqrt{\sum_{f=1}^F V_{fh}}}.$$

동일 가계 내 sib(s)들의 일배체형들이 서로 독립이라고 가정하자. 질병 관련 표지자와 질병 간의 연쇄 및 연관성을 검정하기 위해 z_h에 기초한 두 통계량을 제안하고자 한다.

$$T_1 = \max_{i=1,\ldots,l}|z_{h_i}|,$$
$$T_2 = \frac{l-1}{l}\sum_{i=1}^{l} z_{h_i}^2.$$

T_1은 Mantel-Haenszel 형태의 통계량인 z_h의 절대값 중에서 가장 큰 값을 갖는 일배체형에만 의존하는 통계량인데 반해, T_2는 서로 다른 l개의 일배체형들이 서로 독립이라고 가정하고 각 일배체형에 대응하는 z_h의 제곱을 합한 통계량이다. Spielman과 Ewens (1998)과 Monks 등 (1998)는 각각 단일 표지자가 이대립인자와 다대립인자(multi-allele)를 가질 때 표지자와 질병 간의 연쇄 및 연관성 검정을 위한 통계량을 제안했는데 위에서 제안한 두 통계량은 그들의 통계량을 일배체형으로 확장한 것이다. 따라서 Spielman과 Ewens (1998)과 Monks 등 (1998)처럼 T_1에 기초한 검정은 Monte Carlo 순열 검정(permutation test)으로 가능하고 (Boehnke와 Langefeld, 1998; Monks 등, 1998; Spielman과 Ewens, 1998), T_2에 기초한 검정은 순열 검정이나 근사적으로 자유도 $(l-1)$를 갖는 카이제곱 검정으로 가능하다.

2.3. 일배체형 쌍을 추정할 경우

지금까지는 발단자를 포함하여 모든 sib(s)의 유전자형에 대응하는 일배체형 쌍을 알고 있다고 가정하였다. 그러나 이형접합성 유전자형의 유전자 좌위가 2개 이상 존재하면 그 유전자형에 대응하는 일배체형의 쌍이 불확실하기 때문에 그와 같은 가정은 사실상 만족되지 않는다. 이와 같은 어려움을 극복하기 위해 본 논문에서는 Zhao 등 (2000)과 김진흠 등 (2004, 2005)이 제안한 방법을 도입하여 불확실성이 존재하는 유전자형에 대해서는 확률적 배분을 통해 일배체형 쌍을 결정하고자 한다. 그러므로 r_{fh}와 s_{fh}은 이제 더 이상 유일하게 결정되지 못하고 오직 조건부 확률을 통해서만 추정될 수 있다. 유전자형 G_g에 대응하는 모든 순서화된 일배체형 쌍들의 집합을 $\mathcal{H}_g(g=1,\ldots,k)$ 라고 놓자. 총 l 개의 일배체형에 대한 미지의 빈도를 $f_h(h=h_1,\ldots,h_l)$ 라고 하자. f_h의 추정방법으로 Clark의 알고리즘 (Clark, 1990), EM 알고리즘 (Excoffier와 Slatkin, 1995), Gibbs 샘플링 방법 (Stephens 등, 2001), Partition-ligation 방법 (Niu 등, 2002) 등이 널리 쓰이고 있는데 본 논문에서는 EM 알고리즘 방법을 사용하고자 한다. 일배체형 h의 빈도 f_h에 대한 추정값을 $\hat{f}_h(h=h_1,\ldots,h_l)$ 라고 하자. 무작위 교배(random mating)와 Hardy-Weinberg 평형을 가정하고서

$$D_g = \Pr(G_g|f_h, h=h_1,\ldots,h_l) = \sum_{(s,t)\in\mathcal{H}_g} f_s f_t, \quad g=1,\ldots,k,$$
$$w_{stg} = \Pr(\text{일배체형 쌍}=(s,t)|G_g) = f_s f_t/D_g, \quad s,t=h_1,\ldots,h_l;\ g=1,\ldots,k$$

으로 정의하자. 따라서 O_{fh}, r_{fh}와 s_{fh}에 대한 추정값 $\hat{O}_{fh}, \hat{r}_{fh}$와 $\hat{s}_{fh}$은 조건부 확률 w_{stg}를 통해 아래와 같이 얻어진다.

$$\hat{O}_{fh} = 2\sum_{g=1}^{k} x_{fg}\left\{\sum_{(s,t)\in\mathcal{H}_g}\hat{w}_{stg}I(s=h,t=h)\right\} + \sum_{g=1}^{k} x_{fg}\left[\sum_{(s,t)\in\mathcal{H}_g}\hat{w}_{stg}\{I(s=h,t=m,m\neq h)+I(s=m,t=h,m\neq h)\}\right],$$

$$\hat{r}_{fh} = \sum_{g=1}^{k} t_{fg}\left\{\sum_{(s,t)\in\mathcal{H}_g}\hat{w}_{stg}I(s=h,t=h)\right\},$$

$$\hat{s}_{fh} = \sum_{g=1}^{k} t_{fg}\left[\sum_{(s,t)\in\mathcal{H}_g}\hat{w}_{stg}\{I(s=h,t=m,m\neq h)+I(s=m,t=h,m\neq h)\}\right].$$

단, $\hat{D}_g = \sum_{(s,t)\in\mathcal{H}_g}\hat{f}_s\hat{f}_t$, $\hat{w}_{stg} = \hat{f}_s\hat{f}_t/\hat{D}_g$ 이다. 한편 $E_{fh}, V_{fh}, z_h, T_1, T_2$에 포함된 r_{fh}와 s_{fh}를 각각 $\hat{r}_{fh}$와 $\hat{s}_{fh}$로 치환하고, z_h, T_1, T_2에 포함된 O_{fh}를 $\hat{O}_{fh}$로 치환하여 얻은 것을 $\hat{E}_{fh}, \hat{V}_{fh}, \hat{z}_h$, $\hat{T}_1, \hat{T}_2$ 하자. T_1과 T_2와 달리 두 통계량 $\hat{T}_1$과 $\hat{T}_2$의 점근분포는 $\hat{r}_{fh}$와 $\hat{s}_{fh}$의 복잡성 때문에 쉽게 유도되지 않아 본 논문에서는 순열 검정방법을 써서 두 검정법의 통계적 유의성을 알아보고자 한다. Monks 등 (1998)에 의해 제안된 순열 검정은 아래와 같은 절차에 따라 수행된다.

- 단계 0: 주어진 자료로부터 검정통계량 $\hat{T}_1, \hat{T}_2$를 계산하여 각각 $\hat{T}_{10}, \hat{T}_{20}$ 라고 놓는다.
- 단계 1: 가계별로 N_f^a와 N_f^u를 유지하면서 각 sib(s)의 질병 상태(affected/unaffected)를 무작위로 배치한다.
- 단계 2: 의사 표본(pseudo-sample)으로부터 검정통계량 $\hat{T}_1, \hat{T}_2$를 계산하여 각각 $\hat{T}_{1b}, \hat{T}_{2b}$ 라고 놓고, 이 두 값에 대응하는 $\hat{T}_{10}, \hat{T}_{20}$와 대소관계를 각각 알아본다. 단, $b = 1, \ldots, B$ 이다.
- 단계 3: 단계 1과 단계 2를 B번 반복하고, 실험적 유의확률 값(p-값)은 B번의 반복 중에서 $\hat{T}_{1b}, \hat{T}_{2b}$이 대응하는 $\hat{T}_{10}, \hat{T}_{20}$ 보다 큰 값을 갖는 빈도로 각각 정의한다.

3. 모의실험 연구

모의실험에서는 3개의 표지자 좌위에서 각각 2개의 대립인자 1과 2를 갖는 모형을 가정하고자 한다. 질병 관련 유전자 좌위에서는 이대립인자 D와 d를 갖는다고 가정하자. 이때 가능한 8개의 일배체형을 각각 $h_1 = (1,1,1), h_2 = (1,1,2), h_3 = (1,2,1), h_4 = (1,2,2), h_5 = (2,1,1), h_6 = (2,1,2), h_7 = (2,2,1), h_8 = (2,2,2)$ 라고 하자. 8개의 일배체형 중에서 h_7과 h_8은 질병 관련 대립인자 d를 포함하고 있고, 그 외 나머지 6개의 일배체형은 와일드 대립인자(wild allele) D를 포함하고 있다고 하자. 그러므로 h_7과 h_8은 고위험(high risk) 일배체형, 그 외 나머지 일배체형들은 저위험(low risk) 일배체형으로 구분된다.

표 3.1: 일배체형의 분포와 집단 혼합에 따른 모집단 분포

유형	집단	$(f_{h_1}, f_{h_2}, f_{h_3}, f_{h_4}, f_{h_5}, f_{h_6}, f_{h_7}, f_{h_8})$
I	1	(0.125,0.125,0.125,0.125,0.125,0.125,0.125,0.125)
	2	(0.250,0.000,0.250,0.000,0.250,0.000,0.250,0.000)
II	1	(0.343,0.147,0.147,0.063,0.147,0.063,0.063,0.027)
	2	(0.490,0.000,0.210,0.000,0.210,0.000,0.090,0.000)

제안한 검정통계량이 제1종 오류율을 잘 조절하는지 알아보기 위해 일배체형의 분포와 집단 혼합에 따라 두 가지 유형의 모집단 분포를 표 3.1과 같이 가정하였다. 유형 I에서는 집단 1의 분포를 균일하게 가정한 반면에 유형 II에서는 일배체형에 따라 다르게 가정하였다. 집단 2의 분포는 유형 I, II에 관계없이 세 번째 유전자 좌위에서 유전자 변이가 일어나지 않은 것으로 가정하였다. sib(s)의 수(2명/5명)와 가계의 수(50가구/100가구)의 조합으로 가능한 4가지 모든 경우에 대해 모의실험을 수행하였다. 유형 I, II에 관계 없이 집단 1에 속한 개체가 질병에 노출될 가능성 즉, 기저 위험율을 0.1로 가정한 반면에 집단 2에 속한 개체는 0.2, 0.3 혹은 0.4로 가정하여 집단 1과 집단 2가 서로 다르도록 하였다. 같은 크기의 표본을 1,000번 반복하여 뽑고, 각 표본에 대해 300번의 순열 검정을 통해 실험적 유의확률을 구하였다. 모집단의 유형(I/II)과 집단 2의 기저 위험율(0.2/0.3/0.4)에 따른 검정통계량 $\hat{T}_1$과 $\hat{T}_2$의 실험적 유의수준의 결과를 표 3.2에 나타냈다. 또한 EM 추정량의 정도(precision)를 평가하기 위해 일배체형 쌍을 정확하게 재조합하는 비율(단위:%)을 구하여 표 3.2에 제시하였다 (Rohde와 Fuerst, 2001).

먼저 일배체형 쌍을 정확하게 재조합하는 비율을 살펴보면 표본의 크기보다는 모집단의 분포에 더 의존하는 것으로 나타났다. 이는 3개의 표지자에서 가능한 유전자형이 총 27개로 다양하지 않아 표본의 크기(가계의 수 혹은 sib(s)의 수)가 늘어나도 단지 동일한 유전자형의 개수만 증가할 뿐 추정량의 정도를 개선하는 데는 영향을 주지 못했기 때문이라고 생각한다. 유형 I에서의 비율이 유형 II에서의 비율보다 약간 작은 것은 유형 I의 일배체형 분포가 유형 II보다 상대적으로 더 균등하여 일배체형 쌍의 가능도(likelihood)가 어느 한 쌍으로 치우칠 가능성이 적기 때문이라고 생각한다. 한편 1종 오류율에 대한 추정값의 표준오차가 0.0069이므로 제1종 오류율에 대한 95% 신뢰구간은 대략 (0.037, 0.064) 이다. 표 3.2에 있는 모든 값이 모집단의 유형이나, 가계의 수 그리고 sib(s)의 수에 관계없이 이 구간에 포함되기 때문에 제안한 검정법은 제1종 오류율을 매우 잘 조절하고 있다고 말할 수 있다.

제안한 두 검정통계량 $\hat{T}_1$와 $\hat{T}_2$의 검정력을 비교하기 위해 유전 양식으로는 우성모형(D)과 열성모형(R)을 가정하였다. 우성모형에서는 고위험 일배체형인 h_7과 h_8 중에서 어느 하나만 갖고 있어도 그렇지 않은 개체보다 상대위험(RR)이 증가하지만, 열성모형에서는 h_7과 h_8를 모두 갖고 있을 때만 증가한다. 모의실험에서는 상대위험을 1,2,3,4,5로 가정했는데 그 중에서 특히 RR=1에 대응하는 검정력은 제1종 오류율을 의미한다. 표 3.2에서 살펴보았듯이 제안한 통계량은 집단 혼합으로부터 로버스트한 성질을 갖고 있기 때문에 검정력을 비

표 3.2: 모집단의 유형과 집단 2의 기저 위험율에 따른 검정통계량 $\hat{T}_1$과 $\hat{T}_2$의 유의수준과 일배체형 쌍을 올바르게 재조합하는 비율(%)

		집단 2의	sib(s)의 수					
		기저	2			5		
유형	가계의 수	위험율	비율	$\hat{T}_1$	$\hat{T}_2$	비율	$\hat{T}_1$	$\hat{T}_2$
I	50	0.2	80.0	0.041	0.053	80.4	0.049	0.048
		0.3	80.0	0.054	0.050	80.3	0.054	0.056
		0.4	80.1	0.051	0.051	80.7	0.053	0.050
	100	0.2	80.0	0.055	0.048	80.1	0.047	0.042
		0.3	80.1	0.053	0.058	80.1	0.044	0.043
		0.4	80.0	0.044	0.047	80.2	0.052	0.046
II	50	0.2	85.6	0.043	0.052	85.9	0.050	0.057
		0.3	85.7	0.047	0.050	85.7	0.052	0.056
		0.4	86.0	0.048	0.050	85.8	0.048	0.048
	100	0.2	85.4	0.047	0.045	85.5	0.040	0.050
		0.3	85.4	0.053	0.055	85.7	0.054	0.052
		0.4	85.4	0.046	0.051	85.7	0.050	0.044

교할 때는 유형 I과 II에서 집단 2는 제외하고 집단 1로만 이루어져 있다고 가정하였다. 기저 위험율은 0.1로 가정하였고, 총 200개의 가계에 2명 혹은 5명의 sib(s)가 있다고 가정하였다. 유의수준 5%에서 검정통계량의 실험적 검정력은 같은 크기의 표본을 1,000번 뽑고, 각 표본에 대해 300번의 순열 검정을 통해 얻었다. 유전 양식과 모집단 유형에 따라 변하는 검정통계량의 검정력을 그림 3.1에 나타냈다. 그림 3.1에서 점선과 실선은 각각 검정통계량 $\hat{T}_1$과 $\hat{T}_2$의 검정력을 표시하고, 사각형과 원은 각각 sib(s)의 수가 2명, 5명일 때의 검정력에 해당한다. 유전 양식과 모집단 유형에 따른 4가지 모든 조합에서 공통적으로 발견되는 검정력의 특징은 검정통계량 $\hat{T}_2$가 $\hat{T}_1$보다 우수하고, sib(s)의 수가 많아질수록 검정력이 증가했다. 특히 후자의 경향은 자료 수가 증가함으로써 나타난 결과라고 말할 수 있다. 그러나 2.1절에서 언급한 '조건2'를 만족하는 sib(s)만이 검정력을 증가시키는데 영향을 미치기 때문에 자료 수의 증가($2 \rightarrow 5$)에 비해 검정력의 증가폭은 그다지 크지 않다고 생각한다. 한편 모집단 유형에 따른 검정력은 일배체형의 분포가 상대적으로 더 균일한 유형 I이 유형 II보다 더 큼을 알 수 있었다. 따라서 모집단의 분포가 몇몇 일배체형으로 치우쳐 있을수록 그 만큼 거짓인 귀무가설을 검출하기가 어려워질 수 있다고 생각된다. 유전 양식에 따른 비교에서는 우성모형이 열성모형보다 훨씬 더 우수한 검정력을 보여주었다. 특히 열성모형에서 유형 II는 유형 I보다 검정력이 매우 낮은데 그 이유는 한 발단자가 변이 동형접합자(homozygote)가 될 가능성이 유형 II는 0.006으로 0.047(유형 I)보다 매우 낮기 때문인 것으로 생각된다.

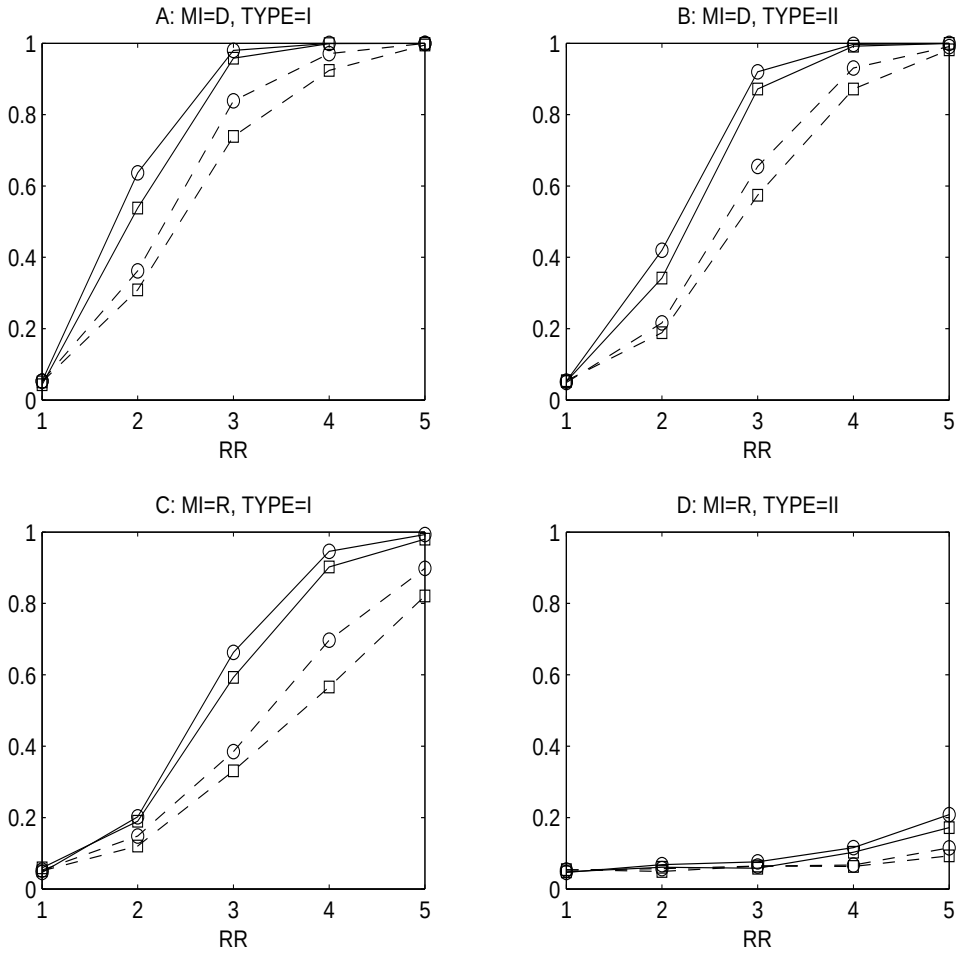

그림 3.1: 유전 양식(MI: D/R)과 모집단 유형(TYPE: I/II)에 따른 검정통계량 $\hat{T}_1$과 $\hat{T}_2$의 검정력 (점선과 실선은 각각 $\hat{T}_1$과 $\hat{T}_2$에 해당하고, 사각형과 원은 각각 sib(s)의 수가 2명, 5명인 경우에 해당함.)

4. 적용 예

연세대학교 심혈관계질환 유전체연구센터에서 수집한 자료 중에서 AGT 유전자 좌위에서 조사한 4개의 단일염기서열다형성(single nucleotide polymorphism; SNP)을 통해 AGT 유전자와 고혈압의 연쇄 및 연관성 분석을 수행하였다. 분석에 사용된 4개의 SNP 중에서 3개의 SNP는 프로모터(promoter) 영역에 존재하고, 나머지 1개는 엑손(exon) 2 영역에 존재한다. 전자는 G-217A, A-20C, G-6A이고, 후자는 M235T 이다. 총 조사된 92개의 가계 중에서 어느 한 sib의 유전자형이 결측되거나 2절에서 언급한 '조건1'을 만족하지 못하는 즉,

표 4.1: 검정통계량 $\hat{T}_1$과 $\hat{T}_2$의 관찰값과 10,000번의 순열 검정을 통해 얻은 실험적 p-값

SNP 개수	SNP(s)	가계수 (유효 가계수)	$\hat{T}_1$ 관찰값	$\hat{T}_1$ p-값	$\hat{T}_2$ 관찰값	$\hat{T}_2$ p-값
4	s_1, s_2, s_3, s_4	26(16)	1.342	0.538	4.382	0.420
3	s_1, s_2, s_3	26(15)	1.342	0.559	2.796	0.475
	s_1, s_2, s_4	28(17)	1.087	0.669	2.337	0.585
	s_1, s_3, s_4	28(15)	1.357	0.370	4.042	0.295
	s_2, s_3, s_4	26(11)	1.342	0.461	3.982	0.312
2	s_1, s_2	35(19)	0.662	0.836	0.621	0.795
	s_1, s_3	28(14)	1.286	0.424	2.400	0.403
	s_1, s_4	31(16)	1.302	0.373	2.460	0.373
	s_2, s_3	26(10)	1.342	0.479	2.300	0.304
	s_2, s_4	28(12)	1.087	0.525	1.859	0.502
	s_3, s_4	28(9)	1.357	0.240	3.124	0.236
1	s_1	38(13)	0.176	1.000	0.031	1.000
	s_2	35(12)	0.494	0.775	0.244	0.775
	s_3	28(7)	1.151	0.334	1.324	0.334
	s_4	44(11)	1.109	0.382	1.230	0.382

모든 sib(s)의 질병 상태(affected/unaffected)가 동일한 가계는 분석에서 제외하였다. 그러나 '조건2'를 만족하지 못하는 가계는 검정력에는 영향을 미치지 않지만 포함하여 분석하였다. 분석에 포함된 SNP의 개수(1개∼4개)에 따라 검정통계량 $\hat{T}_1$과 $\hat{T}_2$의 관찰값을 계산하였고, 10,000번의 순열 검정을 통해 실험적 p-값을 구하여 표 4.1에 나타냈다. 표 4.1에서 SNP의 표기를 단순하게 하기 위해 G-217A는 s_1, A-20C는 s_2, G-6A는 s_3, M235T는 s_4로 바꾸어 나타냈다. 가계수는 분석에 포함하는 SNP의 개수와 종류에 따라 변하며 2절의 조건 1과 2를 모두 만족하는 유효 가계수도 괄호 안에 함께 나타냈다. 수집된 자료에는 sib(s)의 수가 2명인 가계가 많아 유효한 가계가 적게 나왔으며, 그 크기는 7∼16 가계에 이르렀다.

SNP을 1개만 포함하여 분석했을 때 두 통계량의 관찰값은 달라도 검정력이 일치하는 것은 두 통계량이 서로 비례 관계를 만족하기 때문이다. 유의수준 0.05이하의 p-값을 갖는 SNP은 없었지만 4개의 SNP 중에서 가장 유의한 p-값을 갖는 SNP는 G-6A로 나타났다. SNP를 2개만 포함할 때 가능한 조합은 총 6 가지인데 그중에서 가장 유의한 조합은 G-6A와 M235T 이었으며, SNP를 1개만 포함했을 때의 결과와 계보를 이루고 있음을 알 수 있었다. p-값의 변화를 보면 0.334에서 0.236으로 크게 줄어들어 AGT 유전자와 고혈압의 연쇄 및 연관성 분석에서 G-6A만을 고려하는 것보다는 G-6A와 M235T를 함께 고려하는 것이 통계적으로 더 유의함을 알 수 있었다. SNP를 3개 동시에 포함했을 때는 G-6A와 M235T 조합의 계보를 유지하면서 G-217A가 추가로 포함되었지만 p-값의 변화는 0.236에서 0.295로 약간 증가했다. 이는 SNP를 하나 더 고려함으로해서 유효 가계수가 9에서 15로 6만큼 증가

표 4.2: 분석에 포함된 SNP의 개수에 따른 일배체형에 대한 EM 추정량

SNP(s)							
(s_1, s_2, s_3, s_4)		(s_1, s_3, s_4)		(s_3, s_4)		s_3	
일배체형	$\hat{f}_h$	일배체형	$\hat{f}_h$	일배체형	$\hat{f}_h$	일배체형	$\hat{f}_h$
AAAT	-	AAT	-	AT	-	A	0.8281
AAAC	0.1983	AAC	0.2031	AC	0.8281		
AAGT	-	AGT	-	GT	0.1641	G	0.1719
AAGC	-	AGC	-	GC	0.0078		
ACAT	-						
ACAC	-						
ACGT	-						
ACGC	-						
GAAT	-	GAT	-				
GAAC	0.4655	GAC	0.6250				
GAGT	0.1810	GGT	0.1641				
GAGC	0.0086	GGC	0.0078				
GCAT	-						
GCAC	0.1466						
GCGT	-						
GCGC	-						

했지만, 이와 동시에 통계량의 자유도도 3에서 7로 4만큼 증가하였고, 본 자료는 자유도의 증가가 오히려 유효 가계수의 증가에 따른 검정력을 압도했기 때문이다. 마지막으로 4개 모든 SNP를 포함하여 분석하면 유효 가계수는 1개(15→16) 증가하는데 비해 통계량의 자유도는 8만큼(7→15) 증가하기 때문에 p-값이 더 작아지리라고 예상할 수 없다. 표 4.1에서 볼 수 있듯이 G-217A, G-6A, M235T를 포함할 때보다 4개 모든 SNP를 고려했을 때 오히려 통계적 유의성이 떨어짐을(0.295→0.420) 알 수 있었다.

한편 분석에 포함된 SNP의 개수에 따라 가장 통계적으로 유의한 SNP의 조합에 대해서만 일배체형의 EM 추정량을 표 4.2에 나타냈다. SNP의 개수가 1개일 때는 G-6A의 대립인자 A와 G의 분포를 구했는데 모집단에서 A가 차지하는 비율은 G보다 5배 가량 높은 것으로 나타났다. SNP의 개수가 2개일 때는 G-6A와 M235T의 대립인자들의 조합으로 가능한 일배체형 (A,T), (A,C), (G,T), (G,C)의 분포를 구했는데 각각 0%, 83%, 16%, 1%로 나타났으며, (A,C)와 (G,T)이 지배적임을 알 수 있었다. affected sib(s)와 unaffected sib(s)들만의 일배체형 분포는 표 4.2에 따로 표시하지 않았는데 전자는 0%, 79%, 21%, 0%로 나타났고, 후자는 0%, 86%, 12%, 2%로 나타났다. 따라서 고혈압 소인이 있는 사람들은 그렇지 않은 사람들보다 일배체형 (G,T)를 상대적으로 많이 갖고 있는 것으로 생각된다. SNP의 개수가 3개일 때는 (G-217A, G-6A, M235T)의 조합으로 총 8개의 일배체형이 가능한데, 그중

에서 4개의 일배체형 (A,A,C), (G,A,C), (G,G,T), (G,G,C)에만 분포하였다. 마지막으로 모든 SNP를 포함했을 때는 가능한 16개의 일배체형 중에서 오직 5개의 일배체형 (A,A,A,C), (G,A,A,C), (G,A,G,T), (G,A,G,C), (G,C,A,C)에만 집중하여 분포하는 것으로 나타났다.

5. 결론 및 고찰

고혈압, 당뇨 등과 같은 성인병이나 암과 같은 다인성 질병과 여러 유전자의 연쇄 및 연관성 분석은 단일 표지자 대신 일배체형을 바탕으로 하는 연구로 그 중심이 옮겨 가고 있는 추세이다. 이에 발맞추어 본 논문에서는 일배체형에 기초한 sib-TDT 검정법을 제안하였으며 모의실험연구를 통해 제안한 두 검정법 모두 유의수준 5%에서 제1종 오류율을 잘 조절한다는 것과 집단 혼합에 대해 로버스트한 성질을 갖고 있음을 확인할 수 있었다. 또한 우성모형에서는 일배체형의 분포가 이질적일 때보다 동질적일 때 검정력이 더 우수한 것으로 나타났으며, 열성모형에서는 일배체형의 분포에 따라 검정력이 민감하게 변할 수 있음을 알 수 있었다. 한편 표 3.2에 나타나 있지는 않지만 집단 혼합이 있는 모의실험에서 제1종 오류율의 추정값이 명목형 유의수준에 대한 95% 신뢰구간에 포함되지 못하는 경우가 발생하곤 했다. 이를 극복하기 위한 한 방법으로는 순열(permutation) 횟수를 늘려 제안 통계량의 기각역의 값을 더욱 정확하게 추정하는 방법을 고려해 볼 수 있다고 생각한다.

연세대학교 심혈관계질환 유전체연구센터로부터 수집한 자료에 대해 AGT 유전자와 고혈압의 연쇄 및 연관성 분석을 한 결과는 유효 가계수가 적어서(7∼16 가계) 통계적으로 유의한 결과를 얻을 수는 없었다. 그러나 분석에 포함되는 SNP의 개수에 따라 통계적으로 가장 유의한 SNP들이 다음과 같이 계보를 형성하고 있음을 알 수 있었다.

G-6A → G-6A, M235T → G-217A, G-6A, M235T → G-217A, A-20C, G-6A, M235T

위 4 가지 일배체형의 조합 중에서 통계적으로 가장 유의한 것은 (G-6A, M235T)으로 밝혀졌는데, 일배체형에 포함되는 SNP이 많아질수록 표본의 크기가 커져 검정력이 증가하는 측면이 있지만 검정통계량의 자유도도 함께 증가하기 때문에 효율적인 SNP들로 이루어진 일배체형, 소위 tagging-SNP들로 이루어진 일배체형을 형성하는 것이 중요하다는 것을 알 수 있었다.

본 논문에서는 동일 가계 내 sib(s)들의 일배체형들 사이에 존재할 수 있는 상관관계를 고려하지 않고 검정법을 제안했기 때문에 한계점을 갖고 있다고 생각한다. 향후 연구에서는 이런 점을 극복하기 위한 새로운 방법론을 시도하고자 하며, 적은 유효 가계수의 문제는 부-모-자로 이루어진 가계의 자료를 포함하는 통합적인 통계량을 만들어 해결해보고자 한다. 또한 본 논문에서 사용한 EM 알고리즘을 포함하여 일배체형의 추정방법에 따른 제안 통계량의 검정력을 비교해보고자 한다. 이 비교 결과는 실제 자료의 여러 다양한 상황에서 가장 적합한 일배체형의 추정방법을 선택하는 데 도움을 줄 수 있을 것으로 기대한다.

감사의 글

귀중한 임상자료를 제공해주신 연세대학교 심혈관계질환 유전체연구센터에 감사드리

고, 분석 자료의 정리에 조언을 해주신 연세대학교 임상시험센터 강대룡 교수님께 깊이 감사드립니다.

참고문헌

김진흠, 강대룡, 서일, 남정모 (2005). 일배체형에 기초한 고혈압과 ACE 유전자의 연관성 분석, <응용통계연구>, **18**, 297–310.

김진흠, 강대룡, 이윤경, 신선미, 서일, 남정모 (2004). 일배체형에 기초한 연쇄분석의 통계학적 알고리즘의 연구, <예방의학회지>, **37**, 366–372.

Bishop, Y. M., Fienberg, S. E., Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press, Cambridge.

Boehnke, M. and Langefeld, C. D. (1998). Genetic association mapping based on discordant sib pairs: The discordant-alleles test, *The American Journal of Human Genetics*, **62**, 950–961.

Clark, A. G. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations, *Molecular Biology and Evolution*, **7**, 111–122.

Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population, *Molecular Biology and Evolution*, **12**, 921–927.

Monks, S. A., Kaplan, N. L. and Weir, B. S. (1998). A comparative study of sibship tests of linkage and/or association, *The American Journal of Human Genetics*, **63**, 1507–1516.

Niu, T., Qin, Z. S., Xu, X. and Liu, J. S. (2002). Bayesian haplotype inference for multiple linked single nucleotide polymorphisms, *The American Journal of Human Genetics*, **70**, 157–169.

Rohde, K. and Fuerst, R. (2001). Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information, *Human Mutation*, **17**, 289–295.

Spielman, R. S. and Ewens, W. J. (1998). A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test, *The American Journal of Human Genetics*, **62**, 450–458.

Stephens, M., Smith, N. J. and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data, *The American Journal of Human Genetics*, **68**, 978–989.

Zhao, H., Zhang, S., Merkikangas, K. R., Trixler, M., Wildenauer, D. B., Sun, F. and Kidd, K. K. (2000). Transmission/disequilibrium tests using multiple tightly linked markers, *The American Journal of Human Genetics*, **67**, 936–946.

[2007년 9월 접수, 2007년 12월 채택]

Transmission and Disequilibrium Tests Based on Sibship Data*

Jinheum Kim[1] Yangsoo Jang[2]

Abstract

Family-based tests such as the transmission and disequilibrium tests(TDT) have proved to be powerful tools in the search for disease genes. Unlike case-control studies, the tests are not affected by population admixture, which can lead to spurious association of multiple highly linked makers with disease-susceptible genes. Those tests have largely required knowledge of parental marker genotypes. However, parental data are often not available for late-onset diseases. In this article we propose sib-TDTs that overcome this problem by use of marker data from unaffected sib(s) instead of parents. To do this end, we first defined a Mantel-Haenszel-type statistic for each haplotype and then proposed two tests based on this statistic. Simulation studies suggest that the proposed tests are robust to population admixture and are monotone increasing as a relative risk increases irrespective of mode of inheritance. We also illustrated the proposed tests with data adopted from Yonsei Cardiovascular Genome Center

Keywords: Angiotensinogen gene, association, haplotype, linkage, permutation test, transmission and disequilibrium test.

* This work was supported by the Korean Research Foundation Grant funded by the Korean Government (MOEHRD)(KRF-2006-312-C00087)

1) Corresponding author. Associate Professor, Dept. of Applied Statistics, University of Suwon, Gyeonggi -Do 445-743, Korea.
E-mail:jinhkim@suwon.ac.kr

2) Professor, Division of Cardiology, Cardiovascular Genome Center, Yonsei University College of Medicine, Seoul 120-749, Korea.
E-mail: jangys1212@yuhs.ac