예방의학회지 제40권 제2호(2007년 3월)
J Prev Med Public Health 2007;40(2):108-113

특집

유전체 코호트 연구의 주요 통계학적 과제

박소희

국립암센터 암등록역학연구부 암통계연구과

Statistical Issues in Genomic Cohort Studies

Sohee Park

Cancer Biostatistics Branch, Division of Cancer Registration and Epidemiology, National Cancer Center

When conducting large-scale cohort studies, numerous statistical issues arise from the range of study design, data collection, data analysis and interpretation. In genomic cohort studies, these statistical problems become more complicated, which need to be carefully dealt with. Rapid technical advances in genomic studies produce enormous amount of data to be analyzed and traditional statistical methods are no longer sufficient to handle these data. In this paper, we reviewed several important statistical issues that occur frequently in large-scale genomic cohort studies, including measurement error and its relevant correction methods, cost-efficient design strategy for main cohort and validation studies, inflated Type I error, gene-gene and *gene-environment interaction and time-varying hazard* ratios. It is very important to employ appropriate statistical methods in order to make the best use of valuable cohort data and produce valid and reliable study results.

J Prev Med Public Health 2007;40(2):108-113

Key words : Cohort studies, Epidemiologic methods, Validation studies, Research design, *Measurement error*, *Gene-environment* interaction, Proportional hazards models, Genomics

서 론

역학 연구를 성공적으로 수행하기 위해서는 정밀하게 설계된 연구계획이 필수적이며, 연구 설계로부터 데이터 수집, 분석, 결과 해석, 보고서 및 논문 작성에 이르는 전 단계에 걸쳐 발생하는 통계적인 문제들을 올바르게 판단하고 대처하는 일이 매우 중요하다. 특히 대규모 유전체 코호트 연구와 같이 장기간의 연구기간과 추적조사가 필수적인 경우 이러한 통계적 문제를 어떻게 해결하는가에 따라 도출된 연구 결과의 타당성와 신뢰도가 결정될 뿐만 아니라, 자료 수집과 활용 단계에서 많은 자원을 절약할 수도 또는 낭비할 수도 있게 된다. 본 논문에서는 대규모 유전체 코호트 연구에서 발생하는 여러 통계적인 문제점들 중, 특히 질병 발생의 위험요인으로 여겨지는 설명변수의 노출량 측정 시 발생하는 오차로 인한 문제점과 타당성 연구를 통한 그 보정 방법, 생체지표를 이용한 노출량 측정과 반복측정 개수 선정, 다중검정으로 인한 1종오류의 증가, 유전-유전 또는 유전-환경 상호작용 분석법, 시간-의존 위험비 추정 등을 중심으로 고찰하고자 한다.

노출변수 측정에서의 오차와 타당성 연구(Validation study)

1. 측정오차(measurement error)의 영향

측정오차는 거의 모든 노출변수를 측정할 때 발생한다고 볼 수 있다. 노출변수 측정값에 포함된 오차는 연구하고자 하는 노출변수와 결과변수의 상관관계를 설명하는데 있어 비뚤림(bias)을 야기하게 된다. 많은 연구자들이 이러한 측정오차를 간과하는 경향이 있으나, 노출변수를 측정하는데 포함된 오차는 연구결과를 예상치 못 한 방향으로 비뚤리게 할 수 있으므로 많은 주의가 요구된다. 따라서 오차의 크기를 적정하게 추정하고 이를 기반으로 측정오차에 대한 보정을 하는 방법을 고려하는 것이 중요하다.

예를 들어 측정오차는 지방 섭취량과 유방암의 관계를 밝히는 여러 연구에서 일관되지 않은 결과를 보이는 원인으로서 제기된 바 있다. 생태학적 연구, 환자-대조군 연구 또는 개입 연구에서는 유의한 연관성을 보인 지방 섭취량과 유방암 발생의 관계가 대규모 코호트 자료를 이용한 합동분석(pooled analysis) 결과에서는 전혀 유의하지 않은 결과를 보인 것에 대하여 [1], Prentice는 코호트 연구에서 설문지를 사용하여 측정하는 식이 섭취량이 오차를 포함하고 있어 실제로는 연관성이 있는 관계가 그렇지 않은 것처럼 귀무가설 쪽으로 비뚤리기 때문이라고 지적했다 [2]. 식이섬유 섭취량과 대장암 발생의 관계에서도 비슷한 양상이 나타났는데, 환자-대조군 연구에서는 식이섬유 섭취량이 많은

이 연구는 질병관리본부 학술용역사업(2006-347-2400-2440-215)으로 수행된 내용에 근거한 것임.
책임저자 : 박소희 (경기도 고양시 일산동구 마두1동 809번지, 전화 : 031-920-2180 , 팩스 : 031-920-2034, E-mail : shpark@ncc.re.kr)

사람에서 대장암 발생 상대위험도가 낮아진다고 보고된 반면, 코호트 연구 또는 합동분석 연구에서는 유의한 연관성을 보이지 않았다 [3,4].

측정오차는 크게 임의오차(random error)와 계통오차(systematic error)로 구분된다. 임의오차만을 포함하는 측정치는 임의로 참값에 비하여 크거나 작은 오차를 포함하며 따라서 개인 내에서 여러 번에 걸친 측정을 반복할수록 그 평균값이 참값에 가까워지는 경우를 말한다. 반면 계통오차는 아무리 여러 번 측정을 반복하여도 그 평균값이 참값에 근접하지 않고 오히려 참값에서 멀어진 다른 값을 갖게 되므로 이 오차에 대한 적절한 이해와 통제가 없을 경우 연구결과에 심각한 비뚤림을 초래할 수 있다. 나아가 단변량 분석을 넘어선 다중분석을 할 경우 한 개 이상의 설명변수가 측정오차를 포함하고 있을 때 비뚤림은 더욱 복잡한 형태의 예측할 수 없는 방향으로 생기게 되며, 이 때 오차를 포함한 변수들 간 상관성이 존재할 때에는 잔차 교란효과(residual confounding effect)가 생기게 되어 더욱 왜곡된 결과를 얻을 수 있다 [5]. 간혹 연구자들이 설명변수에서 측정오차가 포함된 경우에는 비뚤림이 무조건 귀무가설 쪽으로 생긴다고 잘못 알고 있는 경우가 있는데, 위에서 살펴본 바와 같이 오차가 어떤 형태로 존재하는가에 따라서 비뚤림은 여러 방향으로 생길 수 있다. 또한 측정오차가 버크슨(Berkson) 오차 모형을 따르는 경우에는, 노출변수와 종속변수를 설명하는 관계의 점추정치에서는 비뚤림이 생기지 않는다 [6].

2. 타당성 연구(validation study)와 참고치(reference measurement)를 이용한 측정오차 보정법

먼저 한 개인에서 측정하고자 하는 노출변수의 참값을 X_i로 표시하고 일반적인 방법을 이용하여 측정된 값을 Q_i로 표시했을 때, 이 측정값이 임의오차를 포함한다고 가정하면 고전적인 측정오차모형(classical measurement error model)은 다음과 같이 표현된다.

$$Q_i = X_i + \varepsilon_i,\ i=1,\cdots,n,\ \text{그리고}\ \varepsilon_i \sim N(0, \sigma^2_\varepsilon)$$

예를 들어 식이 섭취량과 질병 발생의 관계를 규명하는 영양역학 연구에서 대상자의 장기간에 걸친 식습관을 평가하고자 할 때 반정량적 식품섭취빈도설문지(semi-quantitative food frequency questionnaire, 이하 빈도설문지라 함)가 널리 이용되는데, 이 빈도설문지에 의해 추정되는 식이 섭취량을 오차를 포함한 Q_i라고 볼 수 있다. 이 때 오차에 해당하는 ε_i의 크기를 추정하고 이에 대한 적절한 보정을 하기 위해서 타당성 연구(validation study) 또는 보정연구(calibration study)를 수행할 수 있다.

측정오차의 크기를 추정하기 위해서는 타당성 연구 대상자에서 노출변수를 흔히 두 가지 방법으로 측정하게 되는데, 첫째는 원 코호트 연구의 모든 대상자에서 사용되는 오차를 포함하는 일반적인 측정방법이고, 둘째는 황금기준(gold standard)에 해당하는 참고치(reference measurement)를 수집하는 방법으로서 대개의 경우 첫 번째 방법에 비해 비용과 노력이 훨씬 많이 든다. 예로 영양역학 연구에서 식이 섭취량을 측정하기 위하여 일반적인 방법으로 빈도설문지를 사용하는 반면, 참고치로는 24시간 회상법, 식사기록지 등을 사용하는 것을 들 수 있다.

이러한 타당성 연구는 원 코호트 연구에 포함되는 내부 타당성 연구(internal validation study)의 형태를 취하기도 하고, 원 코호트에 포함되지 않는 다른 대상자에서 수행되는 외부 타당성 연구(external validation study)가 될 수도 있다. 하버드 보건대학원에서 진행되어온 Nurses' Health Study에서는 빈도설문지를 사용함으로 야기되는 식이 섭취량에서의 측정오차 크기를 타당성 연구 대상자의 자료에서 회귀보정방법(regression calibration method)을 적용함으로써 추정한 뒤, 전체 코호트 연구 자료에서 추정되는 질병 발생 상대위험도를 이러한 측정 오차에 대하여 보정하여 비뚤림을 통제하는 방법을 제안하였다.

이는 타당성 연구 대상자에서 황금기준값(gold standard)에 해당하는 참고치와 일반 측정방법을 이용한 관측치 간의 회귀방정식을 적합한 뒤, 여기에서 얻어지는 기대치($E(X|Q)$)를 전체 코호트 대상자에서 오차를 포함한 관측치 대신에 대치하여 사용하는 방법이다. 그러나 이 때 대치값을 사용함으로 인하여 모수 추정치에 대한 분산값이 실제보다 작게 추정되는 문제가 발생하게 된다. Rosner 등은 델타방법을 이용하여 분산값을 보정함으로써 모수 추정치의 유의성을 검정하는 방법을 제안하였다 [7,8].

3. 코호트 연구(main cohort study)와 타당성 연구(validation study)의 대상자수 선정

측정오차의 범위 추정과 이에 대한 보정을 위해 타당성 연구(validation study) 또는 보정 연구(calibration study)를 계획하는 경우, 전체 코호트 연구에서 몇 명을 타당성 연구에 포함시킬 것인가를 결정해야 한다. 타당성 연구에서 사용되는 참고치가 황금기준값이라는 가정 하에는 고전적 측정오차 모형을 이용하여 대상자수를 선정하는 방법을 Spiegelman과 Gray가 제안한 바 있다 [9].

그러나 오차를 포함한 측정치의 타당성 검증을 위하여 사용되는 참고치도 대부분의 경우 참값을 추정할 뿐 참값 그 자체가 될 수는 없다. 예로 영양자료의 경우 24시간 회상법, 식사일기, 3일 식이기록지 등을 이용한 측정값도 어느 정도의 오차를 포함한다고 가정된다. 따라서 통상적으로는 이러한 측정값을 여러 번에 걸쳐 얻은 뒤 이들의 평균치를 참값의 불편추정값(unbiased estimator)으로 사용하게 되는데, 이 때 반복측정을 몇 번 하는 것이 최적의 연구 설계방법이 되는가는 또 하나의 중요한 통계적인 문제이다. 최적의 연구를 설계하는 방법으로는 정해진 유의수준과 통계적인 검정력을 유지하면서, 측정오차 추정에 필요한 타당성 연구 대상자수(n)와 반복측정 개수(m)를 최소화하고, 노출변수와 질병 발생 여부의 관계를 설명하기 위해 필요한 코호트 연구 대상자수(N)를 동시에 최소화하는 알고리즘을 적용할 수 있다. Park과 Stram은 전체 코호트 크기(N), 타당성 연구 크기(n), 참조치의 반복측정

개수(m)를 동시에 최적화하는 비용-효율적인 연구설계방법을 제안하였다. 최적의 연구설계방법을 결정짓는 항목에는 기저질병발생율(baseline disease rate), 주 코호트 연구에서 검정하고자 하는 노출변수 변화량에 따른 상대위험도의 크기(hypothesized log relative risk), 질병 발생 여부 정보 수집 비용, 황금기준값 추정을 위한 참고측정치의 측정비용, 노출변수 측정값에서 임의오차와 계통오차의 크기, 참값과 측정값간의 상관성 등이 포함된다 [10].

최적의 연구 설계 방법에서 타당성 연구 크기(n)가 전체 코호트 크기(N)에 비해 차지하는 비율을 측정오차의 크기의 함수로 나타내 보면, 측정오차가 크면 클수록, 즉 측정치와 황금기준값 간의 상관성이 작으면 작을수록, 그리고 검정하고자 하는 상대위험도가 클수록 전체 코호트에 비한 타당성 연구의 상대적인 크기가 증가하는 것을 볼 수 있다 (Figure 1). 또한 타당성 연구에서 참고치를 최적의 개수만큼 반복 측정한 경우와 필요 이상으로 반복 측정한 경우 전체 연구의 자료 수집 비용을 살펴보면 필요 이상으로 여러 번 참고치를 반복 측정하는 경우에는 노출변수와 질병 발생 여부의 관계를 추정하는 데에 있어 통계적 검정력의 향상은 없이 정보 수집을 위한 전체 비용이 증가하는 것을 볼 수 있다 (Figure 2). 특히 드문 질병의 경우 이러한 연구 설계는 최적의 연구 설계 방법에 비하여 상대적으로 큰 비용 손실을 초래할 수 있음을 보인다. 따라서 타당성 연구를 수행할 때에는 전체 코호트 연구의 크기와 함께 사전에 주의 깊은 연구계획이 필요할 것이다.

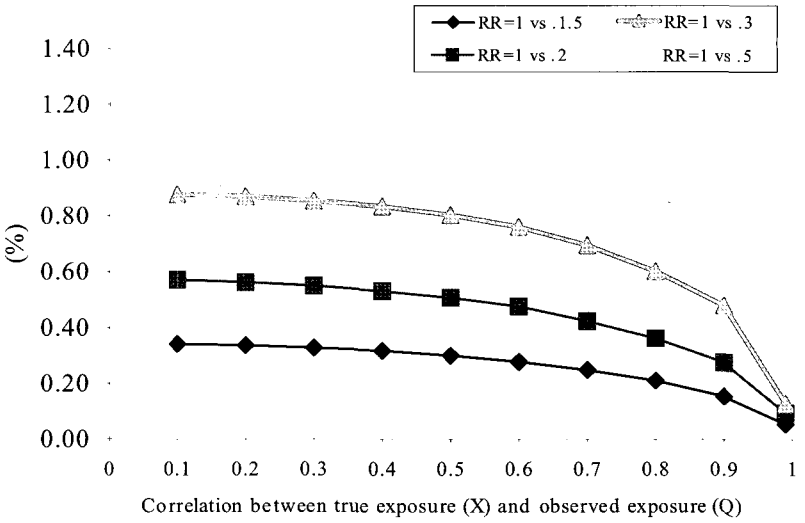

Figure 1. The fraction (in %) of validation study size in the entire study as a function of correlation between true exposure and observed exposure, under a general error model.

* Cost to obtain a reference measure (Z) is assumed to be 50 times the cost to obtain a conventional measure (Q); Power=90%; Significance level=5%; Baseline disease rate=0.5%
(Source: Park and Stram, 2002)

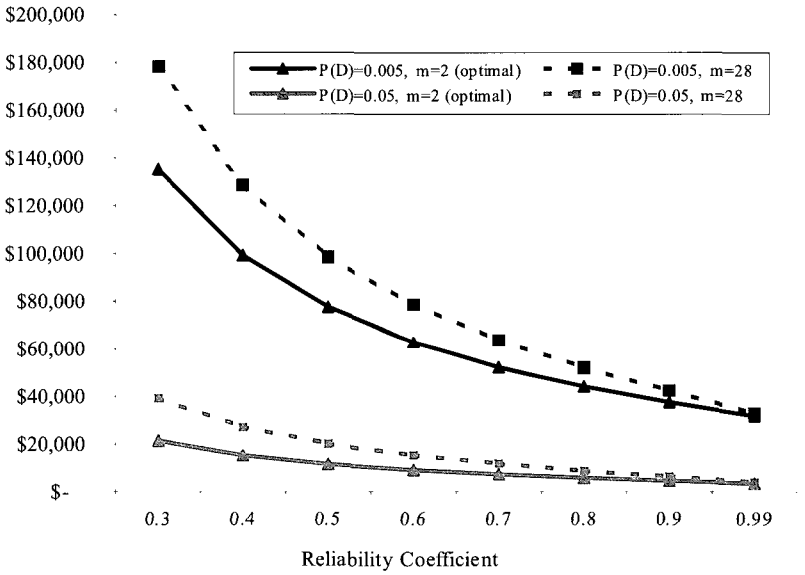

Figure 2. Total cost for obtaining exposure and disease information as a function of the reliability coefficient, when the replicates of the imperfect reference measurement is used.

* P(D): Baseline disease probability; m = number of replicates of reference measures; Assumption: cost to obtain disease information=$1; conventional measure (Q) = $1; reference measure (Z) = $50; Power=90%; Significance level=5%; Hypothesized relative risk = 1.5
(Source: Park and Stram, 2002)

4. 생체지표(biomarker)를 이용한 노출량 추정과 반복측정 방법

측정오차 문제는 흔히 황금기준으로 여겨지는 생체지표(biomarker)를 통해 노출량을 측정하는 경우에도 완전히 없어지는 것은 아니다. 많은 연구자들이 설문지를 통해 얻은 노출량은 측정오차를 포함하지만 생체지표를 이용한 측정치는 참값이라고 믿는 경향이 있으나 이는 사실이 아니다. 생체지표 값을 이용할 때에도 그 값의 반감기가 얼마인지, 개개인의 동역학이 어떠한 영향을 미치는지 등, 그 값의 속성에 따라 주의 깊게 사용되어야 할 것이다. 예로 프탈레이트(phthalates) 노출량의 생체지표로 사용되는 요 대사물(urinary metabolites)의 경우, 개인 내 반복 측정된 값들에서 유의한 일별 변화량과 월별 변화량이 관찰되었으며 [11] 실제로 이러한 양상은 많은 생체지표 값에서 존재한다. 따라서 생체지표 값을 어느 한 시점에서만 수집한다면, 이는 개인의 노출량을 추정함에 있어 큰 오차를 포함하게 되며 질병과 노출량의 관계를 연구하고자 할 때 비뚤림을 야기할 수 있다. 따라서 생체지표를 측정할 때 반복측정을 몇 차례 수행해야 하는가 뿐 아니라 어느 시점에서 측

정할 것인가를 결정하는 일도 매우 중요한 통계적인 과제이며 이는 전체 코호트 연구 결과에서 보고자 하는 유효크기(effect size) 등을 함께 고려하여 결정되어야 할 것이다. 극히 소수의 부분적 대상자에서만 생체지표 값이 반복 측정되는 경우 이 값들 간의 복잡한 공변량 구조가 생길 수 있는데, 예를 들어 측정치에서 일별, 월별, 계절별 변화량 등이 존재하는 경우이다. 이를 고려할 수 있는 측정오차 보정법이 추정방정식(estimating equation)의 틀에서 잠재변수모형(latent variable model)을 이용하여 개발되고 있다 [12].

또한 이러한 노출량 측정값이 임의오차뿐 아니라 계통오차를 포함하는 경우에는 개체특정적 임의효과(subject-specific random effect)를 오차모형에 적용하는 방법이 제안되었으나, 개인별 오차의 크기가 연령, 비만도 등의 개인 성향에 따라 다른 경우에는 이 방법도 제한적이 되므로, 생체지표를 사용한 노출량 측정을 계획할 때에도 이러한 통계적 이슈들을 고려해야 할 것이다 [13].

유전체 코호트 자료 분석 시 발생하는 쟁점

1. 1종 오류의 증가와 이를 통제하는 통계적인 방법

하나의 연구 내에서 여러 검정을 동시에 수행하는 경우 각 연구가설 당 1종 오류의 허용범위를 기존의 5%로 적용하는 경우 false positive 결과를 도출하게 될 확률이 크게 증가한다. 다중검정에서의 1종 오류 증가에 대한 통계적인 논의는 오래 전부터 되어 왔고, 임상 및 역학 연구 등에서는 이러한 1종오류 보정에 치우치다 보면 중요한 연구 결과를 간과할 수 있으므로 연구결과의 중요도를 p값으로 판단하지 말아야 한다는 의견들도 있었다 [14,15]. 그러나 대규모의 유전체 코호트 연구에서 수많은 가설을 검정하는 경우에 증가되는 1종 오류의 문제는 중요하게 다뤄져야 할 것이다.

다중검정 수행으로 인한 문제점을 보정하는 통계기법으로는 먼저 가장 쉽고 간단하게 사용할 수 있는 본페로니 수정(Bonferroni correction)방법이 있다. 이는 FWER (family-wise error rate)을 통제함으로써 다중검정의 문제점을 해결하는 고전적인 방법으로 간단히 적용할 수 있다는 장점이 있으나 조사하는 개수가 증가할수록 유의수준이 급격히 감소하는 단점을 지니고 있으며 검사의 수가 많은 경우에는 규칙이 너무 엄격하여 통계적 검정력이 매우 낮아지게 된다.

이에 반해 Benjamini와 Hochberg는 전체적인 FWER (family-wise error rate)을 통제하는 대신 FDR (false discovery rate)을 통제하는 방법을 제안하였다 [16]. 적용하는 방법은 다음과 같이 비교적 간단하다. 먼저 연구에서 동시에 검정된 가설의 개수가 k라고 할 때, 구해진 k개의 p값을 가장 작은 값으로부터 순서대로 나열한다 (i=1, …, k). 그리고 전체 연구에서 통제할 1종오류를 α라 할 때, i번째로 작은 p값을 $i\alpha/k$의 유의수준에서 기각한다. 이 경우 가장 작은 p값은 본페로니 수정방법에서 사용하는 α/k의 유의수준에서 기각되고, 그 다음으로 작은 p값부터는 본페로니 수정방법보다는 조금 더 관대한 규칙을 적용하게 된다고 볼 수 있다. 그러나 FDR 방법은 다중으로 수행되는 여러 번의 검정이 서로 독립적(independent)이라는 가정을 가지고 있어, 예를 들어 대규모 유전체 코호트 연구에서 한 유전자 내 여러 개의 SNP을 동시에 분석하는 경우와 같이 서로 상관성이 있는 많은 검정을 동시에 수행하는 경우에는 적합하지 않을 수 있다.

이 밖에도 다중검정에서 몇 개의 유의한 결과를 얻었을 경우 FPRP (false positive report probability)를 이용하여 그 결과가 true positive가 아닐 확률을 계산함으로써, 얻어진 유의한 결과가 주목할 만한지 결정하는 방법이 제안되었다 [17]. 수많은 검정을 동시에 수행하는 대규모 유전체 코호트 연구 등에 사용될 수 있는 최적의 1종 오류 통제법은 앞으로도 더 연구되어야 할 주요 통계적 문제로 남아있다.

2. 유전-유전 또는 유전-환경 상호작용 분석방법

질병 발생과의 관계에 있어 실제로는 많은 유전자가 관여될 수 있으며, 각 유전자들 간의 상호작용이 질병 발생 여부와 연관성을 가질 수 있다. 이러한 유전-유전 상호작용(gene-gene interaction)을 통계유전학에서는 에피스타시스(epistasis)라고 지칭하며, 다수의 유전변이에 의한 효과가 각각의 유전변이에 의한 효과를 단순히 합한 (또는 곱한) 것과는 다른 경우를 말한다. 마찬가지로 유전변이에 의한 효과는 환경노출의 차이에 따라 다를 수 있는데, 이 경우 유전-환경 상호작용에 대한 분석이 필요하다. 실제로 수많은 유전자와 환경 요인에서 각각의 주 효과(main effect) 뿐 아니라 상호작용의 효과를 검정하기 위해서는 먼저 충분한 대상자수가 확보되어야 한다. 단순히 두개의 요인을 고려하는 경우에서도, 요인들 간 상호작용을 검정하기 위하여 필요한 대상자수는 각 요인의 주 효과만을 검정할 때에 비하여 적어도 4배가 된다고 알려져 있다 [18]. 검정하려는 유전 또는 환경 요인의 개수가 늘어남에 따라 필요한 대상자수도 증가하므로 수많은 요인에 대한 주 효과와 상호작용을 검정할 때에는 충분한 통계적 검정력을 확보할 수 있는지 살펴보아야 할 뿐 아니라, 적절한 통계분석 기법을 사용하여야 한다.

유전-유전 또는 유전-환경 상호작용을 검정하기 위해서 가장 쉽게 사용할 수 있는 방법은 일반적으로 사용되는 로지스틱 회귀모형, 포아송 모형, 콕스 비례위험 모형 등에서 각 요인변수의 곱으로 나타나는 유전-유전 또는 유전-환경적 상호작용 변수를 설명변수로 모형에 포함하여 검정하는 방법이 있다. 그러나 이 방법에서는 검정하고자 하는 요인들의 개수와 범주가 증가함에 따라 각 쎌(cell) 내의 자료의 불충분, 점근성(asymptotics)의 부적합 등으로 인한 모형 적용 과정에서의 모수 추정에 문제가 생기게 된다. 또한 각 요인 변수간의 다중공선성(multicolinearity) 문제도 피할 수 없다. 이러한 경우는 '차원의 저주(curse of dimensionality)' 라 일컬어지며, 많은 통계학자들이 이 문제를 해결하기 위한 다양한 방법론을 연구하고 있다. 그 중에는 Richie등이 제안한 MDR (multifactor-dimensionality reduction) 방법이 있는데, 비

모수적이며 유전 모형의 형태에 제한을 받지 않고 다차원의 변수들를 일차원으로 줄여나가 검정하는 방법으로써 유전체 코호트 연구에서 도출되는 코호트 내 환자-대조군 자료 분석 시 사용될 수 있다 [19-21]. 또한 최근에는 고차원의 유전-유전, 유전-환경적 상호작용을 효과적으로 연구하기 위하여 일반화선형모형(generalized linear model)과 나무모형(tree model)의 장점을 이용한 PLTR (partially linear tree-based regression) 방법 등이 제안되었다 [22].

3. 시간-의존 위험비율 (time-varying hazard ratio)

코호트 등의 추적연구에서 질병 발생까지의 시간을 분석하는 방법으로 생존분석이 있다. 특히 준모수적(semi-parametric) 방법을 사용하여 생존시간의 분포에 대한 가정을 필요로 하지 않는 콕스 비례위험모형(Cox proportional hazards model)은 생존자료의 분석에서 가장 자주 사용되는 모형이다. 그러나 이 비례위험 가정이 위배되는 경우에는 콕스 모형에서 추정되는 위험비만으로는 그룹 간의 올바른 비교를 할 수 없다. 이 때에는 비례위험 가정이 위배되는 그룹 변수에 대한 층화 콕스 모형(stratified Cox regression model)을 적합할 수도 있으나, 그 그룹 변수가 생존시간에 미치는 직접적인 영향을 연구하고자 할 때에는 사용할 수 없다. 따라서 위험비 대신 각 그룹 별 누적위험도(cumulative hazards) 또는 누적 질병 발생률(cumulative incidence rates)의 비(ratio)를 사용하여 그룹간의 차이를 추정하는 방법이 제안되기도 하였다. 무작위적 임상시험에 참여한 68,000여 명과 관찰연구에 참여한 93,000 여 명의 폐경 여성들을 대상으로 한 대규모 추적 연구인 Women's Health Initiative study의 연구진에서 최근 발표한 논문에서도 이러한 여러 통계적 문제점들이 논의된 바 있다 [23]. 또한 단지 위험비로서 노출변수와 질병 발생 여부의 관계를 설명하고자 하는 범위를 넘어서 절대위험모형(absolute risk model)을 고려할 수 있다. 근래에 개발되어 사용 범위가 넓어지고 있는 경험적-베이즈(Empirical-Bayes) 방법, 계층모형(hierarchical model) 또는 다수준모형(multilevel model) 등의 적용도 필요하다 [24].

결 론

이상으로 대규모 유전체 코호트 연구를 수행할 때 생길 수 있는 몇 가지 통계적인 주요 문제점들을 살펴보았다. 전반적인 코호트 연구를 수행하는 데에 있어 발생하는 통계적 문제들은 앞서 기술한 내용 외에도 여러 가지가 있으나 지면 관계상 본 논문에서 이들을 모두 다루지는 못 하였다. 유전체 연관 연구에서의 대상자수 산출 문제는 예방의학회지 본 호 "유전체 연관 연구에서의 검정력 및 연구대상수 계산 고찰"에서 Park과 Kim이 따로 고찰하였다. 유전체 코호트에서 수행되는 SNP, haplotype, proteomics 등의 자료를 이용한 연구는 이 분야의 급속한 기술적인 발전과 함께 고전적인 통계적 방법론으로는 해결되지 않는 새로운 차원의 과제들을 생성하고 있으며, 많은 통계학자들이 이에 대한 방법론 개발을 위해 힘쓰고 있다. 향후 우리나라 유전체 코호트 연구에서 귀중하게 수집될 자료들에 대해 고도의 통계적 기법을 적절히 적용하고 여러 문제에 대한 통계적 방법론을 개발하는 일은 매우 중요하며, 타당성 높은 연구결과를 도출하는 데에 있어 필수적이라 여겨진다.

참고문헌

1. Hunter DJ, Spiegelman D, Adami HO, Beeson L, van den Brandt PA, Folsom AR, Fraser GE, Goldbohm RA, Graham S, Howe GR, Kushi LH, Marshall JR, McDermott A, Miller AB, Speizer FE, Wolk A, Yaun SS, Willett W. Cohort studies of fat intake and the risk of breast cancer-a pooled analysis. *N Engl J Med* 1996; 334(6): 356-361
2. Prentice RL. Measurement error and results from analytic epidemiology: Dietary fat and breast cancer. *J Natl Cancer Inst* 1996; 88(23): 1738-1747
3. Fuchs CS, Giovannucci EL, Colditz GA, Hunter DJ, Stampfer MJ, Rosner B, Speizer FE, Willett WC. Dietary fiber and the risk of colorectal cancer and adenoma in women. *N Engl J Med* 1999; 340(3): 169-176
4. Park Y, Hunter DJ, Spiegelman D, Bergkvist L, Berrino F, van den Brandt PA, Buring JE, Colditz GA, Freudenheim JL, Fuchs CS, Giovannucci E, Goldbohm RA, Graham S, Harnack L, Hartman AM, Jacobs DR Jr, Kato I, Krogh V, Leitzmann MF, McCullough ML, Miller AB, Pietinen P, Rohan TE, Schatzkin A, Willett WC, Wolk A, Zeleniuch-Jacquotte A, Zhang SM, Smith-Warner SA. Dietary fiber intake and risk of colorectal cancer: A pooled analysis of prospective cohort studies. *JAMA* 2005; 294(22): 2849-2857
5. Fraser GE, Stram DO. Regression calibration in studies with correlated variables measured with error. *Am J Epidemiol* 2001; 154(9): 836-844
6. Thomas D. New techniques for the analysis of cohort studies. *Epidemiol Rev* 1998; 20(1): 122-134
7. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error. *Am J Epidemiol* 1990; 132(4): 734-745
8. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat Med* 1989; 8(9): 1051-1069
9. Spiegelman D, Gray R. Cost-efficient study designs for binary response data with Gaussian covariate measurement error. *Biometrics* 1991; 47(3): 851-869
10. Park S, Stram DO. Cost-efficient design of main cohort and calibration studies where one or more exposure variables are measured with errors. *Proc Joint Stat Meet* Aug 11-15; New York, NY: 2002. p. 2611-2616
11. Hauser R, Meeker JD, Park S, Silva MJ, Calafat AM. Temporal variability of urinary phthalate metabolite levels in men of reproductive age. *Environ Health Perspect* 2004; 112(17): 1734-1740
12. Park S, Ryan LM, Meeker JD, Hauser R. A latent model for measurement error correction using replicate data. *Proc Int Biom Soc meet* Mar 20-23; Austin, TX: 2005. p. 273
13. Carroll RJ, Ruppert D, Stefanski LA, Measurement Error in Nonlinear Models. New York: Chapman and Hall; 1995. p. 141-164
14. Rothman K. No adjustments are needed for multiple comparisons. *Epidemiology* 1990; 1(1): 43-46
15. Weinberg CR. It's time to rehabilitate the p-value. *Epidemiology* 2001; 12(3): 288-290
16. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful

approach to multiple testing. *J R Stat Soc (Ser B)* 1995; 57(1): 289-300
17. Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *J Natl Cancer Inst* 2004; 96(6): 434-442
18. Breslow NE, Day NE. Statistical methods in cancer research. Volume II--The design and analysis of cohort studies. *IARC Sci Publ* 1987; (82): 1-406
19. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 2001; 69(1): 138-147
20. Motsinger AA, Ritchie MD. Multifactor dimensionality reduction: an analysis strategy for modelling and detecting gene-gene interactions in human genetics and pharmacogenomics studies. *Hum Genomics* 2006; 2(5): 318-328
21. Ritchie MD, Hahn LW, Moore JH. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol* 2003; 24(2): 150-157
22. Chen J, Yu K, Hsing A, Therneau TM. A partially linear tree-based regression model for assessing complex joint gene-gene and gene-environment effects. *Genet Epidemiol* 2007. Epub 2007 Jan 31
23. Prentice RL, Pettinger M, Anderson GL. Statistical issues arising in the Women's Health Initiative. *Biometrics* 2005; 61(4): 899-911
24. Carlin B, Louis TA. Bayes and Empirical-Bayes Methods for Data Analysis. 2nd ed. New York: Chapman and Hall; 2000. p. 57-85