◇ 원저 ◇

의학연구논문의 방법론 및 통계처리기법의 타당성 평가를 위한 점검표 개발

서울대학교 의과대학 예방의학교실

안윤옥 · 이형기

= Abstract =

DEVELOPMENT OF A CHECKLIST FOR ASSESSING THE METHODOLOGICAL AND STATISTICAL VALIDITY OF MEDICAL ARTICLES

Yoon-Ok Ahn, M. D., Hyeong-Ki Lee, M. D.

*Department of Preventive Medicine, Seoul National University College of Medicine,
Seoul, Korea*

It is necessary for medical readers or reviewers to assess critically the methodological and statistical validity of medical articles before accepting their results or conclusion. The authors developed a validity-assessing checklist consisting of 21 items. Among them 14 items for methodological validity included the followings: clear statement of research hypothesis or specific aims, suitable focus, definition of study population(or subjects), eligibility criteria, exclusion criteria, appropriateness of samples, description of methods in detail, desnclosure of accuracy, description of reliability, presence of control, susceptability bias, performance bias, detection bias, transfer bias. The last 5 items are applicable only to analytic study. And 7 errors to statistical validity: incomplete description of basic data, statistical test perfomed yet not defined, incomplete description of power or confidence interval, inadequate description of measures of central tendency or dispersion, incorrect analysis, multiplicity on hypothesis testing, unwarranted conclusion. The first 3 items are 'errors of omission', and the others are 'errors of commission'. The authors suggest the checklist be very helpful, but not perfect. Critical mind is needed, which enables someone to distinguish minor errors from major fallacies.

Key Words: validity assessment, checklist, medical articles

1. 머리말 - 점검표 개발의 배경

오늘날 의학분야에서 전문잡지를 통한 정보의 교환은 가히 폭발적이다.¹⁾ 질병의 새로운 기전이나 원인, 약제나 치료법의 우월성에 대해 모순된 결론을 내리고 있는 수많은 문헌들을 접하며 의학자나 임상 의사들은 그 결과를 인용 또는 적용하기에 앞서 반드시 연구논문의 방법론적 타당성 여부를 평가해야 하게끔 되었다.^{2), 3)} 즉, 연구주제가 아무리 참신하고 연구결과 또는 결론이 대단한 발견으로 이어진다고 해도, 연구주제를 구체화하고 연구의 결론을 유도하기 위해 사용한 연구방법론 및 통계처리기법의 타당성이 의심된다면 연구 전체가 무효화 될 수도 있다.⁴⁾

연구방법론 및 통계처리기법의 측면에서 의학연구논문의 타당성을 평가하기 위해 많은 점검표(checklist)들이 고안, 발표되었다.^{3), 5)-16)} 타 학문영역에서도 유사한 평가작업들이 실시된 바 있다. Dawkins¹⁷⁾는 다중비교, Delucchi¹⁸⁾는 카이자승검정법을 사용하는 과정에서 나타난 오류들을 언급하였으며, 국내에서는 윤기중 등¹⁹⁾과 최종후 등²⁰⁾이 독자적으로 개발한 점검표를 사용하여 학술논문에서 발견되는 방법론 및 통계처리기법의 오류들을 분석-평가하기도 했다. 그러나 이들의 평가방법 및 점검표 항목은 주로 통계학자의 입장을 반영한 것이므로, 이를 그대로 의학연구논문에 적용하기에는 적절하지 않은 점이 많이 있다.

모든 의학연구논문에 적용할 수 있는 방법론적 타당성의 평가기준을 정하는 것이 쉬운 일은 아니다.²¹⁾ Crocker²²⁾는 의학연구논문을 읽을 때 ①연구목적, ②현재의 지식체계에 대한 관련성, ③측정방법의 타당도 및 신뢰도, ④표본의 크기 및 추출방법, ⑤연구구조가 갖는 내재적 결함, ⑥자료분석 결과, ⑦연구성적의 의의(implication) 등을 비판적인 입장에서 점검해야 한다고 하였다. 한편 Esdaile 등²³⁾은 인과

관계의 추론을 목적으로 시행된 관찰연구의 타당성을 평가하는 기준을 제시하였고 Horwitz & Feinstein²⁴⁾도 이와 유사한 기준을 마련한 바 있다. 한편 Evans & Pollock^{12), 13)}, Ambroz²⁵⁾, DerSimonian 등²⁶⁾, Pocock²⁷⁾, Simon & Wittes²⁸⁾, Thacker²⁹⁾는 임상시험의 연구결과를 문헌에 보고할 때 반드시 포함시켜야 할 항목 및 그 기준들을 제시하였다. 특히 DerSimonian 등²⁶⁾은 임상시험의 타당성을 평가하는 점검표 개발에서 반드시 고려해야 할 사항으로 '중요성', '보편성', '용이성(容易性)'의 세가지를 들었는데, 이러한 지침은 임상시험 뿐만 아니라 일반적인 의학연구논문의 타당성을 평가하는 점검표 개발에도 동일하게 적용될 수 있다.

통계처리기법의 타당성을 평가하는 점검표 개발에 크게 두가지 측면에서의 접근이 가능하다. 즉, 적용한 개별 통계기법별로 독립된 점검표를 작성하여 평가하는 것^{19), 30)}과 사용한 기법에 관계없이 전체적인 차원에서 각 기법의 적용 여부에 대한 타당성을 평가하는 방법^{5), 20), 31)}이 있다. 일반적으로 전자의 방법은 개별 통계기법이 올바르게 시행되었는지에만 관심을 갖는다. 따라서 이러한 방법에 의할 경우, 적절한 분석초점이 선택되었으며, 가장 적합한 기법을 활용하고 있는지에 대한 평가가 자칫 소홀해 질 수 있다.

본 연구는, 의학연구논문을 읽거나 실제로 연구를 시행하는 이들이 대상 논문의 방법론 및 통계처리기법의 타당성을 검토할 때 지침이 될 수 있는 점검표의 개발을 목적으로 시행되었다. 먼저 연구방법론의 타당성에 대해서는 중요성 및 보편성의 측면에서 분석적 연구와 기술적 연구에 함께 적용할 수 있는 점검표 항목을 정하고, 이어 분석적 연구에서 정당한 비교가 가능하도록 하기 위해 추가로 요구되는 기준들을 마련하였다. 또한 '분명하게 정의된(explicitly defined)' 기준과, 대한의학협회지에 기발표된 논문 중에서 점검표를 적용하

여 평가한 결과들을 제시하여 독자들이 점검표를 사용하고자 할 때 참고가 되도록 하였다. 한편 통계처리방법의 타당성에 대해서는 사용한 개별 기법의 종류에 관계없이 일관된 적용이 가능한 점검표를 마련하였는데, 명백히 정의된 기준을 제시하고 구체적 평가결과들을 예시한 것은 방법론적 타당성 평가의 경우와 같다.

2. 방법론적 타당성을 평가하기 위한 점검표

1) 연구가설이나 특수목표의 명확한 진술 (clear statement of research hypothesis or specific aims)

분석적 연구에서는 연구가설, 기술적 연구에서는 특수목표가 명료하게 진술되어야 한다. 명료하게 정립된 정확한 문제에 대해 어렵 대답을 하는 것이 불명료한 문제에 대해 정확한 대답을 하는 것보다 낫다.³²⁾ 좋은 가설은 ①틀 또는 그 이상의 변인(變因)사이에 존재하는 관계를 나타내도록 표현되어야 하며, ②실제적인 검증이 가능해야 한다.³³⁾ 연구가설이나 특수목표는 서론부에서 일단 기술되지만 연구의 결론부에서 상응(相應)하는 추상화된 (abstracted) 진술로 표현될 때 비로소 의미를 갖게 된다. 따라서 비록 서론부에서 연구가설이나 특수목표를 기술하였다고 해도 결론부에서 “다음과 같은 결과를 얻었다”라는 표현으로 연구성적만을 부여하고, 이러한 성적으로부터 유도해 낼 수 있는 결론을 분명히 기술하지 않음은 잘못이다.³⁴⁾

2) 적절성(suitability) 및 실천적 명시화(operational identification)

연구자가 측정 또는 관찰한 것이 실제로 알고자 의도했던 속성이나 실체(entity)였다면 이는 매우 ‘적절한’ 측정이다. 요컨대 적절성이란 측정하고자 의도했던 것을 실제로 측정했느냐 하는 표면타당도(face validity) 내지 내용타당

도(content validity)를 의미한다. 특히 분석적 연구의 경우에 연구가설이 실행가설(acting hypothesis)로 대치, 구현되는 과정에서 초기상태, 주동인(主動因) 및 비교동인, 결과변수를 적절하게 선택하는 것이 매우 중요하다.^{35), 36)}

연구가설의 증명 적절하지 못한 초기조건을 설정한 예로, 임신이 Prolactin을 분비하는 뇌하수체선종에 미치는 영향을 조사하기 위해 정상임부를 대조군으로 선택한 것(Γ X X 등, 1985)을 들 수 있다. 그러나 정상임부를 대조군으로 하면 오히려 Prolactin분비선종이 임신에 미치는 영향을 밝혀낼 수 있을 뿐이다. 임신이 Prolactin분비선종에 미치는 영향을 분석적으로 평가하려면 Prolactin 분비선종을 가진 여성으로 임신 상태에 있지 않은 환자를 대조군으로 선택해야 한다. 또한 증상이나 기타 변수에 대한 고려없이 전도속도의 호전만으로 신경병증의 치료 효과를 판정한 것(H X X 등, 1987)은 적절하지 못한 결과변수의 선택이라고 할 수 있다.

한편 적절한 연구초점을 결정하는 과정은 필연적으로 ‘실천적 명시화(operational identification)’라는 단계를 요구하게 된다. 즉, 연구가 실제로 수행되기 위해서는 각각의 요소 및 연구의 매 과정이 구체적으로 정의되어야 함을 의미한다. 기술적 연구나 분석적 연구를 막론하고 실천적 명시화는 다음의 두 단계에서 이루어져야 한다.

(1) 표본의 선택(sampling)

연구의 결론을 적용하고자 하는 대상은 어떤 집단인가? (모집단의 정의) 어떤 환자나 대상들이 선택되었으며, 기준은 무엇인가? (適格性基準) 여러가지 이유들로 대상에서 제외되는 환자들의 기준은 무엇인가? (除外基準)¹⁶⁾ 선택된 연구 대상이 연구의 결론을 적용하려고 하는 집단(target population)을 충분히 대표할 수 있는가? (代表性)

그러나 인간을 대상으로 한 의학연구에서는

대표성을 얻기 위해 무작위추출이나 이에 준하는 추출방법을 적용하기가 매우 어렵다.^{14), 37)}

³⁸⁾ 따라서 표본의 선택 및 제외기준에 대한 기술이 명백해야 하고, 대표성을 유지하기 위해 실시한 방법이나 그 한계점에 대한 언급이 있어야 한다.^{30), 39)}

“본 연구의 목적은 건강한 한국인 흡연자들에서 — 최대 호기 유량곡선에서 계측할 수 있는”(ス X X, 1986)은 연구의 모집단을 잘 기술한 예이다. 그러나 “항결핵제를 투여한 환자 중에서 조기 탈락자를 제외하고 약제투여 중 정기적 추구검사가 가능했고 — 190명을 대상으로(入 X X, 1987)”는 적격성 기준 및 제외 기준을 명확히 기술하지 않은 예이다. 즉, 항결핵제를 투여받은 총 환자수, 여러가지 이유로 대상에서 탈락된 환자수를 명기하지 않은 채, 190명을 대상으로 했다고만 밝히고 있어 대상의 선택과정에 대한 신뢰를 어렵게 하고 있으며 연구결과를 실제로 적용하고자 하는 독자들에게 지침을 마련해 주고 있지 못하다. 요컨대 1900명에서 190명만을 대상으로 선택한 것과 200명 중 190명을 선택한 것은 그 의미가 매우 다르다.

대표성의 한계에 대한 연구자들의 인식부재는, “국내 십이지장궤양의 임상적 보고의 대부분이 소수의 환자를 대상으로 하였기 때문에 통계적 고찰에 다소의 문제점이 있지 않을까 사료된다”(ㄱ X X 등, 1986)라고 기술하면서, 저자 등의 연구는 2,171예를 대상으로 했기 때문에 대표성이라는 측면에서 더 우월하다는 주장을 하는 것에서 엿볼 수 있다. 2,171예가 과연 많은 수인가하는 문제와 ‘통계적 고찰’이 무엇을 의미하는가에 관한 문제는 일단 논의 하더라도, 이것은 추정의 정밀도(精密度)와 대표성을 혼돈하고 있는 증거이다. 연구대상의 숫자가 아무리 많아도 전수조사를 했거나 표본추출의 방법이 무작위추출 또는 이에 준한 확률추출법(probability sample)이 아니라면 여전히 비뚤림이 개입되는 것을 피할 수 없으며

대표성에 대한 신뢰를 보장할 수 없기 때문이다.⁴⁰⁾

(2) 측정(measurement)

연구논문의 방법에는 ①어떤 변수에 대해 어떤 방법을 사용하여 어떻게 자료와 정보를 얻게 되었는지에 대한 상세한 기술, ②일련의 측정방법의 타당도(또는 정확도, validity or accuracy)나 신뢰도(reliability)에 대한 언급이 있어야 한다.

“ethanol 투여 후 얼굴이 붉어지는 군, 붉어지지 않는 군으로 나누었다”(ㄱ X X 등, 1985)는 안면홍조의 여부를 구분한 방법에 대해 명확한 언급을 생략함으로써 연구의 재현성을 저해하고 있다. 한편, “이들 155명의 환자를 대상으로 하여 신뢰 검정을 필하였으며 그 결과는 신뢰 K치가 0.66—0.83이었다”(ㄴ X X 등, 1986)와 같이 측정의 신뢰를 보증하기 위해 실시된 방법을 기술하거나, 설문지 응답의 신뢰성을 밝힌 예(ス X X 등, 1986)는 타당성 기준을 충족시키는 것이된다.

3) 유사성(similarity) 또는 비교성의 확보

유사성 또는 비교성의 확보란 문제는 분석적 연구에서만 중요한 의미를 갖는다. 대조군의 설정은 비교분석을 위한 필수 대전제이다.³⁹⁾

⁴¹⁾ 따라서 요로결석환자 300명에서 475회의 체외충격파쇄석술을 대조처치 없이 시행한 후 “체외충격파쇄석술은 요로결석의 치료에 가장 안전하고 용이하면서 효과적인 방법임을 확인하였다(ㄱ X X 등, 1987)”라는 결론을 내린 것은 타당하지 않다.

한편 공정한 비교를 하기 위해 대조군의 설정과 함께 추가적으로 다음과 같은 4단계에서의 비뚤림을 방지하려는 노력이 중요하다.¹⁴⁾ 즉, 실험군과 대조군의 제 요소가 비교의 각 단계에서 충분히 동일하지 않으면 비뚤림(bias)이 일어나며 이는 연구의 타당성을 저해하는 주 이유가 된다.

(1) 감수성비폴립(susceptibility bias) 또는 선택비폴립(selection bias)

비교하려는 양 집단에 대해 결과사건에 영향을 미치는 예후적 상태(prognostic expectation)가 초기상태에서 같지 않다면 감수성비폴립이 일어난다. 확률할당(무작위할당)이나 층화분석(stratified analysis)은 이러한 비폴립을 최소화하기 위한 방법이다.⁴²⁾ 그러나 대부분의 의학연구에서는 임상시험이나 몇몇 실험실 위주의 연구를 제외하고 확률할당이 사실상 불가능하다. 따라서 '실험'의 형태를 띤 연구논문에서는 ①확률할당이 이루어져야 하고, ②그 방법에 대한 구체적 기술이 있어야 한다. 조사연구에서는 ①중요하다고 생각되는 임의의 요소들이 초기상태에서 비교하려는 양 집단에 대해 동일한지를 구체적 자료의 제시, t-검정, 카이제곱검정 등의 방법을 통해 검토하거나, ②이러한 요소에 대한 층화분석이 이루어져야 한다. 연구대상인 쥐 120마리를 정상식이군(70예)과 Retinyl Palmitate 추가 투여군(50예)의 두 군으로 나누면서 어떤 방법으로 양군의 동질성을 확보할 수 있었는지에 대한 언급을 생략한 것(B X X 등, 1984)은 감수성비폴립의 오류를 범한 예가 된다.

(2) 시행비폴립(performance bias)

비교하려는 주동인(처치방법등)이나 비교동인, 또는 둘 이상의 처치가 동일한 숙련도(proficiency)로 시행되지 않았을 때 시행비폴립이 일어나게 된다. 연구방법에는 약용량, 순응도, 보조동인, 시행자의 숙련도, 처치가 시행된 환경 등의 동일성 내지 유사성을 보증하는 언급이 있어야 한다. 두가지 수술수기를 비교하면서, 동일 내지는 유사한 숙련도를 가진 시술자에 의해 행해졌다는 것과 마취조건 등이 동일하였다는 것을 언급하지 않은 것(츠 X X 등, 1983)이나, 대조군에 위약(placebo)을 투여하지 않은 것(Γ X X, 1986)은 시행비폴립의 예이다.

(3) 발견비폴립(detection bias)

임의의 동인이나 처치의 시행 이후, 결과사건의 발생 여부를 발견하는 과정이 비교하려는 양집단에서 동일하지 않으면 발견 비폴립이 일어난다. 각종 '눈가림법(blind method)'은 이러한 발견비폴립을 최소화하려는 시도이다. "pap smear와 조직생검은 따로따로 판독하여 판독자로 하여금 조직생검의 결과를 모르는 상태에서 pap smear를 판독케 하였다"(Γ X X 등, 1987)는 기술은 타당한 예가 된다.

(4) 탈락비폴립(drop-out bias)

경시적(經時的, longitudinal) 연구의 경우 추적관찰에서 탈락되는 사례(drop-out)가 많아지거나, 외과적 치료의 우월성을 입증하기 위해 수술 후 다시 병원을 찾은 사례만을 대상으로 했다면 전이비폴립이 일어난다. 이를 막기 위해서는 drop-out이 총 대상의 15% 이하가 되게끔 하거나, drop-out된 환자들의 특성에 대한 조사가 병행되어야 한다. 120마리의 쥐 중 생존한 94마리만을 분석대상으로 한 것(B X X 등, 1984)은 탈락비폴립의 한 예가 된다.

3. 통계처리기법의 타당성을 평가하기 위한 점검표

많은 저자들이 의학연구논문에서 흔히 사용되는 통계처리기법에 대해, 이들을 어떻게 선택하고 올바르게 적용할 것인지를 논의하였다. 9), 21), 36), 37), 43)~64) 특히 Gore 등⁷⁾은 통계처리의 잘못을 '생략의 잘못(errors of omission)'과 '시행일 잘못(errors of commission)'으로 크게 나누었는데, 전자는 시행된 분석과정이나 자료 등이 연구자에 의해 충분히 보고되지 않음으로써 발생하는 오류이고, 후자는 통계처리를 오용함으로써 나타나는 오류이다. 본 연구에서는 Gore 등의 분류를 근간으로 세부사항을 수정, 보충하여 다음과 같은 기준을 마련하였다.

1) 생략의 잘못(Errors of Omission)

(1) 기초자료에 대한 불충분한 기술(incomplete description of basic data)

단순히 통계적 유의성만을 언급하는 것보다 연구결과의 양적 표현인 기술통계량(descriptive statistics)을 제시함으로써 장래의 연구를 위한 더 많은 시사점들을 얻을 수 있다.^{27), 37), 50), 65)} 얻어진 자료는 보통 평균(mean), 중앙값(median), 최빈치(mode)와 같이 중심경향(대표치)을 나타내는 측도(measures of central tendency)와 범위(range), 표준편차, inner percentile range, semiinterquartile range 같은 산포측도(measures of dispersion)로 함께 표현되어야 한다. 이들은 상보적(相補的)이며 어느 한 쪽이 다른 쪽을 대신하지 못하므로, 하나가 생략되거나 모두 생략되면 안된다. 뿐만 아니라 $a \pm b$ 의 형태로써 자료의 기술통계량을 제시할 때는 a, b가 의미하는 바를 각각 밝혀야 한다.⁶⁶⁾

(2) 사용된 통계처리기법에 대한 불충분한 기술(statistical test performed yet not defined)

통계적 분석 방법은 상세하게 기술되어야 한다.^{5), 7), 30), 65)} 흔히 적용되는 방법이 아니라면 구체적 내용을 자세히 기술하거나 최소한 참고문헌이라도 밝혀야 한다. '통계 package를 이용하였다'라고만 언급하거나 각 분석 예에 대해 분석방법, 검정통계량(test statistics), 유의수준 중 어느 하나라도 생략되면 '사용된 통계처리기법에 대한 불충분한 기술의 오류'를 범한 것이다. 또한 카이제곱통계량은 여러가지 귀무가설의 검정에 이용될 수 있으므로 — 적합도검정, 동일성 검정, 독립성검정 모두에 카이제곱검정법이 이용될 수 있다 —, 검정하고자 하는 귀무가설을 명시하지 않았을 때에도 동일한 오류를 범한 것이 된다.³⁰⁾

사용한 통계처리기법을 밝히지 않은 채 “이렇게 하여 얻어진 모든 검사치는 통계학적으로

로 처리하였다”(Γ X X등, 1981)라고만 언급한 것은 잘못이다. 비슷한 예로 “GPT치도 투약전 188.9 ± 89.9 에서 투약 후에는 80.8 ± 47.9 로 하강하여 통계학적으로 매우 유의($p < 0.005$)하였다”(日 X X등, 1980)라는 기술은 어떠한 분석기법을 적용했으며, 무엇이 유의한지에 대한 언급을 생략하여 동일한 오류를 범하고 있다.

(3) 검정력이나 신뢰구간에 대한 불충분한 기술(incomplete description of power or confidence interval)

임의의 처치가 기존의 처치 또는 다른 대조처치에 비해 차이가 존재한다는 것을 증명하지 못한 경우 — 보통 negative study라고 한다 —, 이것이 진짜 차이가 없기 때문인지, 아니면 실제로 존재하는 차이를 밝혀내지 못할 정도로 낮은 검정력 때문인지를 밝혀야 한다. 즉, 검정력분석(power analysis)를 실시함으로써, 연구논문을 읽고 이를 적용하려는 독자에게 판단의 근거를 마련해 주어야 한다.^{67), 68)} 낮은 검정력 때문이었다면 ‘확실한 결론에 이르지 못하는 결과(inconclusive results)’이므로 이것을 근거로 ‘차이가 없다’라고 단정짓는 것은 옳지 않다.^{69), 70)} 따라서 연구자는 연구시작 이전의 어느 정도의 차이를 임상적으로 유의하다고 할 것인지를 정하고, 검정력을 고려하여 필요한 표본의 수를 계산해야 한다.^{25), 43), 56), 71)}

또한 연구결과를 읽고 실제로 이를 적용하려는 독자는 단순히 통계학적으로 의미있다 또는 없다고 결정하는 것보다도 양군에서 관찰된 결과의 차이가 얼마나 큰지를 추정하는 것에 더욱 관심을 가질 수 있다.^{72), 73)} 신뢰구간의 제시는 표본의 자료로부터 모수가 위치할 범위를 추정함으로써 이러한 목적에 잘 부합하며, 연구결과의 보고에 신뢰구간을 포함하도록 요구하는 의학잡지의 수가 최근 늘고 있다.

⁷⁸⁾⁻⁷⁸⁾

신뢰구간을 잘 기술한 예로 “The Maentel-

Haenszel estimate of odds ratio was 3.5 and the 95% confidence limits of the odds ratio was 2.3-4.7”(맹광호, 1982)를 들 수 있다.

2) 시행의 잘못(Errors of Commission)

(1) 중심경향(대표치) 또는 산포도에 대한 부적절한 기술(inadequate description of measures of central tendency or dispersion)

자료의 중심경향을 나타내는 측도 선정에는 자료의 특성 및 척도(scale)가 고려되어야 한다. 명목수준의 자료에는 최빈치, 서열척도로 측정된 자료이거나 치우친 분포(skewed distribution)의 자료에는 중앙값이 평균보다 적절한 측도이다.^{30), 43), 49)} 또한 자료의 산포도도 분포를 충분히 고려하여 선정할 필요가 있다. 일반적으로 많이 사용되는 표준편차는 자료가 명백한 비정규성을 보일 때 적절한 산포측도가 되지 못한다.^{58), 79)} 따라서 “환아 18예의 요증 neopterin치는 $2,067.1 \pm 3,260.4$ (mean $\pm$ 1 SD)였고”(ㄱ X X, 1987)와 같은 표현은 연구논문의 원자료(原資料, outlier)들이 존재하며, 오른쪽으로 치우친 분포임을 고려하지 않은 채 습관적으로 표준편차를 산포측도로 제시한 예가 된다.

한편 자료의 산포를 표시하는 측도로서 표준오차(standard error)를 사용함도 잘못이다. 표준오차가 모집단의 임의의 모수(대개는 평균)를 추정할 때 이용되는 추론통계량(inferential statistics)이지 결코 기술통계량이 아니기 때문이다.⁸⁰⁾ 따라서 “본 연구결과는 mean $\pm$ SE로 표시하였다”와 같은 표현(ㅇ X X 등, 1981)은 오류이다. 흔히 이러한 잘못을 범하게 되는 이유는 표준오차가 표준편차나 기타 산포를 나타내는 측도들에 비해서 작은 값을 취함으로써 자료의 정밀도 또는 신뢰정도를 좋아보이게 하기 때문이다.⁸¹⁾

(2) 온당하지 않은 분석방법(incorrect analysis)

sis)

‘온당하지 않다’는 것은 다음의 세 가지 측면에서 오류를 범했음을 의미한다.

첫째, 부적절한 통계학적 ‘robustness’이다. 통계가들은 임의의 통계처리가 근거하고 있는 가정의 중요성을 robustness란 용어로 기술한다.⁶⁴⁾ 통계학적으로 충분한 robustness를 갖는 처리가 되려면 필요한 가정들로부터 많이 벗어나지 말아야 한다. 일례로 전체 표본수가 20 이하이거나, 40이하이면서 예상되는 어느 한 cell의 기대치가 5이하라면 카이자승검정법을 그대로 적용하는 것이 옳지 않으며 연속성 수정을 하거나 Fisher의 직접확률법을 사용해야 한다.^{18), 43)} t검정, 분산분석, 회귀분석을 시행하는 경우에도 각각 만족시켜야 하는 제 가정들이 있는데⁸²⁾, 일반적으로 가설검정에 필요한 가정을 만족시키지 못할 경우에는 비모수통계기법을 이용하거나 자료의 변환(transformation)을 통해 가정에 근접한 형태로 바꾸어 주어야 한다.⁴³⁾

둘째, 통계학적 독립성에 대한 무시이다. 임의의 통계처리를 적용하려면 각 표본에서의 관찰이 상호독립이어야 하며, 한 표본에서의 관찰은 다른 표본에서의 관찰과 독립이어야 한다. 한 대상에서 수회의 관찰을 실시하고 이를 각각 독립된 관찰처럼 취급하여 처리하거나¹⁷⁾, 쌍을 이룬 관찰에 대해 일표본 t검정(unpaired t-test) 또는 일반적인 카이자승검정법을 적용하는 경우, 그리고 비모수통계기법에서 상응하는 종속표본의 검정법을 적절히 적용하지 못한 경우에 ‘통계학적 독립성에 대한 무시의 오류’를 범한 것이 된다. 동일한 환자에서 nifedipine 투여 전후의 수지온도 회복시간을 비교하면서 paired t-test가 아닌 unpaired t-test를 사용한 것(ㅎ XX 등, 1989)은 독립성에 대한 고려를 하지 못했음을 의미한다.

세째, 선택된 통계처리기법의 적절성(suitability)이다. 즉, 주어진 문제해결에 가장 적합한 분석기법을 적용했는가 하는 것이다. 여기

에는 서열척도 이하로 측정된 변수에 대하여 모수적 통계기법을 적용하는 것처럼, 보다 정밀한 척도수준의 분석에 사용해야 할 통계처리 기법을 적용함으로써 논리의 비약을 초래하는 경우, 등간격척도 이상의 척도로 측정된 두 변수의 상관성을 검증하기 위해 이를 범주화하고 비모수상관계수 등을 계산함으로써 정보의 손실을 가져오는 경우¹⁹⁾가 포함된다. 고장액 주입 후 혈중 K^+ 치의 변화를 매 시간마다 도시하고 각각의 값과 기준 시간(0시간)일 때의 값의 차이를 다중 paired t-test로 검정한 것(日 XX등, 1988)은 통계학적 독립성에 대한 고려는 이루어 졌다고 판단되나, RCA(response curve analysis) 기법¹⁷⁾을 적용했다면 단순한 반응의 차이 외에도 유용한 정보들을 분석해 낼 수 있었을 것이다. 임의의 폭로원과 암발생 또는 암발생의 예방이라고 하는 인과관계를 증명하기 위해 암발생율의 차이(risk difference)로 분석한 것(日 XX등, 1982)은 비(比, relative risk)를 이용했다면 하는 아쉬움이 있다. 소공뇌경색증(lacunar infarct)의 심한 정도(severity, 순위척도)와 전산화단층촬영술에 의해 구한 경색의 용적(volume, 비 척도)간의 상관관계를 구하면서 Pearson의 적률상관계수를 이용한 것(入 XX등, 1989)도, 적어도 양 변수가 등간격척도로 측정되었을 때 Pearson의 적률상관계수를 적용해야 하므로 적절하지 않은 분석이다.

(3) 중복시행한 가설검정(multiplicity on hypothesis testing)

임상연구에서 중복가설의 검정은 흔히 ①수집중인 자료를 반복적으로 분석하는 경우, ②상호연관된 둘 이상의 변수에 대해 두 집단을 단변수적(univariate)으로 반복검정하는 경우, ③둘 이상의 아집단(subgroup) 사이에서 다양한 조합의 동시적 분석을 시행하는 경우(simultaneous multiple comparisons)에 일어날 수 있다.⁸³⁾⁻⁸⁵⁾ 이외에 동일한 자료에 대해 여러가

지 형태의 통계기법을 적용하고 이중 연구자가 임의로 선택한 — 대개는 가장 작은 값의 — p-value만을 진정한 것으로 보는 경우^{31), 86)}도 중복시행한 가설검정의 오류를 범한 것이다. 어떠한 경우든지 문제가 되는 것은 ‘제1종의 오류’로서, harris⁸⁷⁾는 “다중비교의 시행시 실제로 우연에 의한 변이외에는 아무 것도 없는데도 최소한 하나 이상의 검정에서 유의한 결론을 얻게 될 확률(type 1 error in multiple tests)은 시행하는 검정의 수가 증가할수록 급격히 증가한다”라고 하였다.

중복가설의 검정에서 일어날 수 있는 이러한 문제를 해결하기 위해 다양한 통계적 기법들이 소개되었다. 동시적 다중 비교(simultaneous multiple comparisons) 방법은 일단 분산분석 등을 통하여 각 집단이 동일한 모집단에서 추출된 것이 아님(전체적으로 집단간에 차이가 있음)을 증명한 후, 개별 각 집단간의 비교와 관련된 오류율을 0.05와 같은 임의의 수준으로 조정하는 것인데 Bonferroni, Tukey, Den-nett, Sheffe, Duncan 등이 각각 자신의 방법들을 개발하였다.⁸⁴⁾ 상호 관련된 여러 변수에 대하여 두 집단을 비교하는 경우에는 변수들 사이의 상관관계를 고려하면서 동시에 양 집단의 차이를 전체적으로 검정할 수 있는 단일화된 분석 기법(multivariate analysis)이 필요하다.⁸⁸⁾

日 XX등(1983)은 간의 담도암에 대해 다른 치료법(A, B, C)을 시행받은 세 군의 혈청 빌리루빈치를 비교하면서 multiple t-test를 이용해 A와 B, B와 C, C와 A를 비교하였다. 개개의 비교에 대해 유의수준을 0.05로 고정한다고 하면, 3회의 검정 이후 제1종의 오류는 $1 - (1 - 0.05)^3 = 0.14$ 로 증가한다.⁴³⁾ 또한 “통계적인 처리는 student t-test와 아울러 성적자체가 표준편차가 큰 분포를 하였으므로 정규분포를 하지 않을 가능성을 고려하여 non-parametric test인 Wilcoxon-rank sum test를 병행하여 비교하였다”(日 XX등, 1982)는 동일한 자료에 대해 여러가지 통계기법을 적용한 뒤, 그 중의

한 방법에 의해 얻어진 p-value만을 선택하는 중복가설검정의 오류를 범하고 있다.

(4) 부당하게 유도된 결론(unwarranted conclusion)

본래 의도했던 결론에 도달하지 못할까 염려하여 연구자는 종종 잘못된 분석과정이나 연구고안(research design)을 통하여 오도된 결론을 유도할 수 있다.⁸⁹⁾ 여기에는 ①표본에 대한 통계적 처리를 전혀 하지 않은 상태에서 모집단에 대한 결론을 유추하는 경우, ②적절한 통계처리를 실시했으나 해석과정에서 정당화할 수 없는 논리의 비약이 있는 경우, ③잘못된 용어를 사용한 경우 — 상관(correlation)을 비교(comparison), 연관(association), 관계(relation)와 같은 용어와 혼동하거나, 특이도(specificty)와 민감도(sensitivity)를 함께 제시하지 않는 경우, 유병율과 발생율의 혼동, p-value에 대한 잘못된 이해, ‘귀무가설이 기각되었다’라고 하지 않고 ‘대립가설이 수용되었다(accepted)’라고 하는 경우, 통계학적 연관(statistical association)과 인과관계를 동일한 것으로 혼동하는 경우 등이 포함된다 —, ④연구결과를 일반 집단에 외삽하면서 그 한계를 언급하지 않은 경우, ⑤여러 개의 집단을 비교하면서 분산분석만 실시하고 다중비교법을 통해 구체적으로 어떤 집단 사이에 차이가 존재하는지를 밝히지 않은 경우(incomplete analysis) 등은 ‘부당하게 유도된 결론의 오류’를 범한 대표적 경우이다.^{9), 10), 53), 59), 90), 91)}

$r=0$ 이라고 하는 귀무가설에 대한 통계적 검증없이 “— $r<0.5$ 로 전혀 상관관계가 없었다 —”라고 기술한 것(XX등, 1983)은 오류이다.

XX등(1986)은 임위의 항암요법에 대한 반응군 및 비반응군의 중간생존기간(median survival time)을 비교한 뒤, 반응군의 중간생존기간이 더 길다는 사실로부터 임의의 항암요법이 우수한 치료법이라는 결론을 내렸다. 그

러나 반응군과 비반응군은 동일한 항암요법을 받는 집단이 두 아집단(亞集團, subset)으로서, 치료에 대한 특성의 반응(예를 들어 종괴의 감소)에 의해 편의상 구분된 집단일 뿐이다. 따라서 반응군과 비반응군의 중간생존기간을 비교하는 것은 치료효과의 우월성을 평가하기 위한 방법이라기보다는, 치료에 대한 반응 여부를 조기에 예측하게 해주는 환자의 특성 — 이를 ‘예후 지표’라고 한다 — 을 밝히고자 할 때 오히려 더욱 유용한 방법이다. 그러므로 반응군과 우월한 생존특성이 결여된 비반응군을 비교하여 반응군의 중간생존기간이 더 길기 때문에 임의의 치료법이 우수하다고 결론짓는 것은 무리이다.⁹²⁾ 동 저자들이 주장하려는 결론을 입증하기 위해서는 유사한 병기(stage), 유사한 병리조직학적 소견, 유사한 연령, 유사한 정신적 건강상태에 있는 둘 이상의 집단에 대해 상이한 치료법을 적용한 후, 각 군의 중간생존기간을 비교함이 옳다.

XX등(1982)은 보건소에 내원했던 월별 환자동향에 의한 자료로부터 “PPNG의 발생율은 급격히 증가하여 9월에는 입원환자 60명 중 27명(45%)이 PPNG에 의한 감염이었다”라고 기술하였다. 그러나 입원환자 60명 중 27명은 비율(rate)이 아닌 분율(proportion)로, 9월 중 PPNG 감염환자가 차지하는 분율이 다른 달의 PPNG 감염환자의 분율보다 높다는 사실에서 결코 PPNG 발생율이 증가하였다는 결론이 유도되지는 않는다. 즉, 이 경우는 분율과 비율을 혼동함으로써 용어에 대한 잘못된 이해의 오류를 범한 것이다.⁹³⁾

XX등(1989)은 “충남을 중심으로 한 농촌 주민을 대상으로” 담석 성분을 분석하고 “이상에서 한국의 농촌지역에서 가장 호발하는 담석은 콜레스테롤 담석이며”라는 결론을 내렸다. 그러나 이것은 연구결과의 무리한 외삽이다.

불충분한 분석의 오류는 ○ XX등(1988), ○ XX등(1989)에서 발견된다. 이들은 세 군간의

평균치를 비교하면서 분산분석만을 실시하고 다중비교를 생략함으로써 오류를 범하였다.

이상 언급한 각 점검항목을 요약하여 도시하면 표 1과 같다.

4. 맺음말-점검표의 타당도 및 신뢰도, 한계

본 연구에서 개발한 점검표는, 의학 관련 제 논문들의 타당성을 평가하기 위해 고안되었던 각종 문헌들이 평가의 기준으로 이용한 항목들을 통합하여 보완한 것이다. 즉, 먼저 연구방법론 또는 통계처리기법의 타당성을 구성한다고 볼 수 있는 몇 개의 하부개념들 — 예를 들어 적절성, 유사성, 생략의 오류와 같은 것들을 말한다 — 을 선정하고 이어 각 하부개념의 의미를 가장 잘 표현한다고 생각되는 기준 항목들로 점검표의 구체적 내용을 구성하였다. 또한 각 항목에 대해 오류를 범했다고 평가할 수 있는 근거 기준을 명백히 정의하였다. 따라서 이는 Bohrnstedt가 제안한 바⁹⁾, 내용타당도 달성을 위한 제 과정을 따르고 있기 때문에 내용타당도에 관한 한 크게 문제될 것은 없다고 보여진다. 한편 기준관련타당도 중 예시적 타당도를 입증하려면 본 연구의 점검표에 의해 타당하지 않다고 평가된 연구논문의 결론들이, 향후 보다 타당하다고 인정되는 연구나 실제 경험 등을 통해서 잘못된 것으로 밝혀지는 것을 추시해야만 한다. 그러나 현 상황에서는 이를 확인할 방법이 없으며, 향후 본 연구와 동일 또는 유사한 점검표를 이용하여 의학 연구논문의 방법론적 타당성을 평가한 연구들에 의해서 밝혀지리라 생각된다.

다음으로 점검표의 신뢰도에 대한 것인데, 평가의 기준(explicit criteria)을 명백히 정의하고, 재조사법에 의한 관찰자내 또는 관찰자간 신뢰도를 평가함으로써 이를 해결할 수 있겠다. 그러나 평가의 성격상 평가자 개인의 관점 및 평가능력에 따라 영향을 받을 수 밖에 없는 한계가 있는 것도 사실이다. 현재로서는 의

학연구논문의 타당성을 평가하는 기타 연구들에 의해 본 점검표의 신뢰도 검증 작업이 이루어지기를 기대할 수 밖에 없다.

요컨대 이러한 점검표가 만능일 수는 없다. 본 연구에서 개발한 점검표는 자칫 숲과 나무를 분간하지 못하게 할 수도 있다. 즉, 점검표에 의해 사독자나 독자들은 연구의 세세한 오류들을 밝혀낼 수 있으나 중대한 오류와 사소한 잘못을 구별하지 않음으로써 전체적인 시각을 어렵게 할 수도 있다. 사실 일점의 흠도 없이 고안, 시행, 분석되는 연구란 불가능하다. 중요한 것은 어떠한 오류가 연구의 타당성에 치명적인지를 결정하는 것이다. 만일 연구자나 독자들이 이러한 결정을 효율적으로 하지 못한다면 세세한 잘못들이 필요이상으로 확대될 것이며, 연구논문의 주요한 공헌들이 평가절하될 것이다. 따라서 점검표에 의해 임의의 항목이 타당하지 않은 것으로 평가되었을 경우에 이것이 전체 연구를 무효화할 정도로 부적절한지를 고려해야 한다. 또한 분석적 연구에서 비뚤림의 문제가 발생할 경우 이러한 왜곡이 발생하게 된 기전을 추구하여 전체적인 비뚤림의 방향을 따져 보는 것이 필요하다.

참 고 문 헌

1. Schwartz DG: Techniques for accessing the medical literature. *J Allergy Clin Immunol* 82:54 4-50, 1988
2. Lionel NDW, Herxheimer: Assessing reports of therapeutic trials. *Br Med J* 3:637-40, 1970
3. Thorn MD, CC Pulliam, MJ Symons et al: Statistical and research quality of the medical and pharmacy literature. *Am J Hospital Pharma* 42:1077-82, 1985
4. 안윤옥, 이형기: 의학에서의 연구방법론, *한국역학회지* 12:107-14, 1990
5. 안윤옥, 고용린: 자료처리과정에 대한 통

- 계학적 검토 — 일부 의학잡지에 게재된
논문예를 중심으로 예방의학지 6:81—
5, 1973
6. Schoolman HM, JM Becktel, WR Best et al: Statistics in medical research: principles versus practices. *J Lab Clin Med* 71: 357—67, 1968
 7. Gore SM, IG Jones, EC Rytter: Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976. *Br Med J* 8:85—7, 1977
 8. O'Fallon JR, SD Dubey, DS Salsburg et al: Should there be statistical guidelines for medical research papers? *Biometrics* 34: 687—95, 1978
 9. Altman DG: Statistics and ethics in medical research: V. Analysing data. *Br Med J* 281:1473—5, 1980
 10. Sahl J: Validity of epidemiologic investigations. *An Allergy* 51(1 pt 1):63—4, 1983
 11. Felson DT, LA Cupples, RF Meenan: Misuse of statistical methods in Arthritis and Rheumatism. *Arthritis and Rheumatism* 27:1018—22, 1984
 12. Evans M, AV Pollock: A score system for evaluating random control clinical trials of prophylaxis of abdominal surgical wound infection. *Br J Surg* 72:256—60, 1985
 13. Evans M, AV Pollock: The inadequacy of published random control trials of antibacterial prophylaxis in colorectal surgery. *Dis Colon Rectum* 30: 743—6, 1987
 14. Feinstein AR: Clinical epidemiology: The architecture of clinical research. 1st ed. W. B. Saunders Company, Philadelphia, 1985
 15. Chalmers I: Misconduct in medical research [letter]. *Br Med J* 298(6668): 256, 1988
 16. Hebert JR, DR Miller: Methodologic considerations of investigating the diet-cancer link. *Am J Clin Nutr* 47: 1068—77, 1988
 17. Dawkins HC: Multiple comparisons misused: Why so frequently in response-curve studies? *Biometrics* 39:789—90, 1983
 18. Delucchi KL: The use and misuse of chisquare: Lewis and Burke revisited. *Psychol Bull* 94:166—76, 1983
 19. 윤기중, 안윤기, 김병수: 통계의 오용과 효율적 이용에 관한 연구. 산업과 경영 24:3—37, 1987
 20. 최중후, 이재창: 학술논문과 통계적 기법. 고려대학교 통계연구소 통계분석강의 총서 12 자유아카데미, 서울, 1990
 21. Feinstein AR: Clinical biostatistics XLIV. A survey of the research architecture used for publications in general medical journals. *Clin Pharmacol Ther* 24:117—25, 1978
 22. Crocker LM. Linking research to practice: suggestions for reading a research article *Am J Occup Ther* 31:34—9, 1977
 23. Esdaile JM, RI Horwitz: Observational studies of cause-effect relationships: an analysis of methodologic problems as illustrated by the conflicting data for the role of oral contraceptives in the etiology of rheumatoid arthritis. *J Chron Dis* 39: 841—52, 1986
 24. Horwitz RI, AR Feinstein: Methodologic standards and contradicting results in case-control research. *Am J Med* 66:556—64, 1979
 25. Ambroz A, TC Chalmers, H Smith et al:

- Deficiencies of randomized control trials. *Clin Res* 26L:280 A, 1978
26. DerSimonian R, LJ Charette, B McPeck et al: Reporting on methods in clinical trials. *N Engl J Med*. 306:1332-7, 1982
 27. Pocock SJ: Statistics in medicine: current issues in the design and interpretation of clinical trials. *Br Med J* 290:39-42, 1985
 28. Simon R, RE Wittes: Methodologic guidelines for reports of clinical trials.[editorial] *Cancer Treat Reports* 69:1-3, 1985
 29. Thacker SB: Quality of controlled clinical trials. The case of imaging ultrasound in obstetrics: A review. *Br J Obstet Gynecol* 92:437-44, 1985
 30. White SJ: Statistical errors in papers in the British Journal of Psychiatry. *Br J Psychiat* 135:336-42, 1979
 31. Schor S, I Karten: Statistical evaluation of medical journal manuscripts. *J Am Med Assoc* 195:1123-8, 1966
 32. Tukey JW: Where do we go from here? *J Am Statist Assoc* 55:80-93, 1960 In: Meire P Statistics and medical experimentation. *Biometrics* 31:511-29, 1975
 33. Kerlinger FN: Foundations of behavioral research. 3rd ed. CBS College Publishing. NEW Youk, 1986. (고흥화, 김현수, 백영승역, 사회행동과학 연구 방법의 기초, 2판. 성원사, pp. 47-61, 1989)
 34. 신영수, 안윤옥: 의학연구방법론(서울대학교 의과대학 대학원 강의록). pp. 126-31, 1989
 35. 채서일: 사회과학 조사방법론. 초판. 법문사, 서울, pp. 111-94, 1990.
 36. Feinstein AR: Clinical biostatistics XLI. Hard Science, soft data, and the challenges of choosing clinical variables in research. *Clin Pharmacol Ther* 22:485-98, 1977
 37. Feinstein AR: Clinical biostatistics XL. Stochastic significance, consistency, apoposite data, and some other remedies for the intellectual pollutants of statistical vocabulary. *Clin Pharmacol Ther* 22:113-123, 1977
 38. Greenstein LR: Teaching research: An introduction to statistical concepts and research terminology. *Am J Occup Ther* 34:320-7, 1980
 39. Saiger GL: Errors of medical studies. *J Am Med Assoc* 173:678-81, 1960
 40. 김동희, 박중양, 손중권 등: 통계학 — 개념과 제문제. 개정판. 자유아카데미, 서울, pp. 3-50, 1989
 41. Ross OB: Use of controls in medical research. *J. Am Med Assoc* 145:72-5, 1951
 42. Feinstein AR: An additional basic science for clinical medicine: III. The challenges of comparison and measurement. *An Intern Med* 99:805-12, 1983
 43. 안윤옥: 실용 의학 통계론. 초판. 서울대학교 출판부, 서울, 1990
 44. Feinstein AR: Clinical biostatistics. Bailer's laws of data analysis. *Clin Pharmacol Ther* 20:113-9, 1976
 45. Feinstein AR: Clinical biostatistics XXXVII. Demeaned errors, confidence games, nonplused minuses, inefficient coefficients, and other statistical disruptions. *Clin Pharmacol Ther* 20:617-31, 1976
 46. Feinstein AR: Clinical biostatistics XXXIX. The haze of Bayes, the aerial palaces of decision analysis, and the com-

- puterized Ouija board. *Clin Pharmacol Ther* 21:482-96, 1977
47. Feinstein AR: Clinical biostatistics XLV. The purposes and functions of criteria. *Clin Pharmacol Ther* 24:479-92, 1978
 48. Feinstein AR: Clinical biostatistics XLVI. What are the criteria for criteria? *Clin Pharmacol Ther* 25:108-16, 1979
 49. Feinstein AR. Clinical biostatistics L. On choosing a mean and other quantitative indexes to describe the location and dispersion of univariate data. *Clin Pharmacol Ther* 27:120-30, 1980
 50. Feinstein AR: Clinical biostatistics LI. Quantitative significance and statistical indexes for a contrast of two groups. *Clin Pharmacol Ther* 27:567-78, 1980
 51. Feinstein AR: Clinical biostatistics LV. The t test and the basic ethos of parametric statistical inference (Part I). *Clin Pharmacol Ther* 29:548-60, 1981
 52. Feinstein AR: Clinical biostatistics LVI. The t test and the basic ethos of parametric statistical inference (Conclusion). *Clin Pharmacol Ther* 30:133-46, 1981
 53. Feinstein AR, Kramer MS: Clinical biostatistics LII. A primer on quantitative indexes of association. *Clin Pharmacol Ther* 28:130-45, 1980
 54. Fleiss JL: Statistical methods for rates and proportions. 2nd ed. John Wiley & Sons, New York, 1981
 55. Altman DG: Statistics and ethics in medical research: Misuse of statistics is unethical. *Br Med J* 281:1182-4, 1980
 56. Altman DG: Statistics and ethics in medical research: III. How large a sample? *Br Med J* 281:1336-8, 1980
 57. Altman DG: Statistics and ethics in medical research: Collecting and screening data. *Br Med J* 281:1399-1401, 1980
 58. Altman DG: Statistics and ethics in medical research: IV. Presentation of results *Br Med J* 281:1542-4, 1980
 59. Altman DG: Statistics and ethics in medical research: VII. Interpreting results. *Br Med J* 281:1612-4, 1980
 60. Altman DG: Statistics and ethics in medical research: VIII. Improving the quality of statistics in medical journals. *Br Med J* 282:44-7, 1981
 61. Braitman LE: Statistical, clinical, and experimental evidence in randomized controlled trials [editoria]. *Ann Intern Med* 98:407-8, 1983
 62. Glantz SA: Biostatistics: how to detect, correct and prevent errors in the medical literature. *Circulation* 61:1-7, 1980
 63. Kramer MS, AR Feinstein: Clinical biostatistics LIV. The biostatistics of concordance *Clin Pharmacol Ther* 29:1123, 1981
 64. Wallenstein S, CL Zucker, JL Fleiss: Some statistical methods useful in circulation research. *Circulation Res* 47:1-9, 1980
 65. Mosteller F: Problems of omission in communication. *Clin Pharmacol Ther* 25:761-4, 1979
 66. Gardner MJ: Understanding and presenting variation. *Lancet* 230-1(Jan.25), 1975
 67. braitman LE: Statistical power analysis in medical reserach [editorial]. *Ann Intern Med* 99:269-71, 1983
 68. Detsky AS, Sackett DL: When was a 'negative' clinical trial big enough? How many patients you needed depends on

- what you found. *Arch Intern Med* 145: 709-12, 1985
69. Freiman J, TC Chalmers, H Smith et al: The importance of beta, the Type II error and sample size in the design and interpretation of the randomized control trial. *N Eng J Med* 299:690-4, 1978
 70. Young MJ, EA Bresnitz, BL Strom: Sample size nomograms for interpreting negative clinical studies. *Ann Intern Med* 99: 248-51, 1983
 71. Kraemer HC: Sample size: When is enough enough? *Am J Med Sci* 296:360-3, 1988
 72. Braitman LE: Confidence intervals extract clinically useful information from data. *An Intern Med* 108:296-8, 1988
 73. Slud EV, DP Byar, SB Green: A comparison of reflected versus test-based confidence intervals for the median survival time based in censored data. *Biometrics* 40:587-600, 1984
 74. Rothman KJ: A show of confidence [letter]. *N Engl J Med* 299-1362-3, 1978
 75. Rothman KJ: Significance questing [editorial]. *Ann Intern Med* 105-445-7, 1986
 76. Simon R: Confidence intervals for reporting results of clinical trials. *Ann Intern Med* 105:429-35, 1986
 77. Bailar III JC, F Mosteller: Guidelines for statistical reporting in articles for medical journals: amplification and explanation. *Ann Intern Med* 108:266-73, 1988
 78. Gardner MJ, DG Altman: Confidence interval rather than p value: estimation rather than hypothesis testing. *Br Med J* 296:746-50, 1986
 79. Charles HH, JE Buring: Epidemiology in medicine. 1st ed. Little, Brown and Company, Boston, pp. 215-42, 1987
 80. Glantz SA: Primer of biostatistics. McGraw-Hill Book Co. New York, 1981
 81. Brown GW: Standard deviation, standard error. Which 'standard' should we use? *Am J Dis Child* 136:937-41, 1982
 82. 채서일, 김범중: SPSS/PC+를 이용한 통계분석. 초판, 법문사, 서울, 1980
 83. Godfrey K: Comparing the means of several groups. *N Engl J Med* 313:1450-6, 1985
 84. Ware JH, F Mosteller, JA Ingelfinger: P values, in Bailar JC, F Mosteller. (eds). Medical use of statistics. Waltham MA. NEJM Books, 1986, 179-210. In: Ottenbacher KJ. Scientific vs. statistical inference: The problem of multiple contrasts in clinical research. *Am J Med Sci* 295:172-7, 1988
 85. Smith DG, J Clemens, W Crede et al: Impact of multiple comparisons in randomized clinical trials. *Am J Med* 83:545-50, 1987
 86. O'Brien PC, A Shampo: Statistical considerations for performing multiple tests in a single experiment. 4. Performing multiple statistical tests on the same data. *Mayo Clin Proc* 63:1043-5, 1988
 87. Harris RJ: A primer multivariate statistics. John Wiley, New York, 1975, pp 201-77. In: Ottenbacher KJ. Scientific vs. statistical inference: The problem of multiple contrasts in clinical research. *Am J Med Sci* 295:172-7, 1988
 88. Cupples LA, T Heeren, A Schatzkin et al.: Multiple testing of hypotheses in comparing two groups. *Ann Intern Med* 100: 122-9, 1984

89. Ford BL, RD Tortora: The consultant's forum. A consulting aid to sample design. *Biometrics* 34:299-304, 1978
90. Diamond GA, JS Forrester: Clinical trials and statistical verdicts: Probable grounds for appeal. *Ann Intern Med* 98:385-94, 1983
91. Long KJ: Abused statistical language in journal articles. *Am J Med* 83:607-8, 1987
92. Condon RE: Type III error [editorial]. *Arch Surg* 121:877-8, 1986
93. Elandt-Johnson RC: Definition of rates: some remarks on their use and misuse. *Am J Epidemiol* 102:267-71, 1975
94. Bohrnstedt GW: "Reliability and validity assessment in attitude measurement", pp. 80-9, in Gene FS ed., *Attitude Measurement*, Rand McNally, Chicago, 1970

부록:본 논문에 인용된 기발표 원저
(대한의학협회지)

- 고철우 등:Kawasaki씨병 환아에서 요중 Neopterin치의 변화. 30:1353-59, 1987
- 김경희 등:한국 정상인 및 간장질환환자의 간 기능에비력에 관한 연구. 26:923-8, 1983
- 김노경 등:성인 급성 임파성 백혈병의 치료에 관한 연구. 29:1346-56, 1986
- 김명환 등:적외선분광계를 이용한 농준주민의 담석 분석. 32:189-98, 1989
- 김상애 등:임신이 Prolactin 분비뇌하수체선종에 미치는 영향에 대하여. 28:763-72, 1985
- 김성수 등:철분투여에 따른 혈액중의 제 철분 지수와 운동수행능력의 변화 — 남녀 농구선수들을 중심으로. 29:

- 1333-45, 1986
- 김세철 등:요로결석 300예에 대한 EDAP-LT01 체외충격파쇄석술의 경험. 30:1134-45, 1987
- 김송명 등:정상한국인의 혈청태아성암항원 정상치에 대한 연구. 24:63-9, 1981
- 김영수 등:간외담도암의 임상적 관찰. 26:163-71, 1983
- 김중환 등:Penicillinase Producing *N. gonorrhoeae* (PPNG)의 급만연. 25:163-8, 1982
- 김택훈 등:임신초기 유산시 Papanicolaou Smear에 의한 태아조직의 진단법. 30:1349-52, 1987
- 김창덕 등:십이지궤양의 임상적 고찰 — 내시경으로 확진된 2,171예를 중심으로. 29:1311-18, 1986
- 김형진 등:음주후 얼굴이 붉어지는 사람에 있어서 혈중 Ethanol 농도곡선의 분석 및 순환기능의 특성. 28:1009-24, 1985
- 맹광호 등:한국 도시 직장인의 고혈압과 체중에 관한 사례-비교군 관찰연구. 25:939-42, 1982
- 민성길 등:획병에 대한 진단적 연구. 29:653-61, 1986
- 박성수 등:급, 만성간염 및 간경변증에 대한 Protoporphyrin Disodium의 치료경험. 23:49-54, 1980
- 박순태 등:Retinyl Palmitate의 투여가 N-methyl-N'-nitro-N-nitrosoguanidine에 의한 Wistar쥐의 상부위장관종양 발생에 미치는 영향에 관한 실험적 연구. 27:65-78, 1984
- 박영춘 등:Gangliosides가 당뇨병성 말초신경병증환자의 운동신경 전도속도에 미치는 영향에 관한 연구. 30:863-68, 1987
- 박재홍 등:유기산에 의한 혈장감염의 변동에

- 관하여. 31:1213-23, 1988
- 손경선 등:항결핵 단기요법의 성과고찰. 30: 1109-14, 1987
- 손영호 등:소공뇌경색증의 임상적 연구. 32: 757-66, 1989
- 원영천 등:Lorazepam이 이산화탄소 반응곡선에 미치는 영향. 29:409-16, 1986
- 이문호 등:한국형출혈열의 병인론에 관한 연구. 25:325-40, 1982
- 이명덕 등:아플라톡신 B₁, 메주 및 N-Methyl-N'-Nitro-N-Nitrosoguanidine에 의한 Wistar쥐의 위암발생에 관한 실험적 연구. 25:149-62, 1982
- 이병철 등:한국인 당뇨병환자 1,017례의 구획별 신경전도검사에서 나타난 말초 신경손상의 양상. 31:757-70, 1988
- 이홍규 등:한국인 비만형 인슐린 비의존형 당뇨병에서의 지방조직의 지방산 구성. 32:397-406, 1989
- 정만홍 등:남자 흡연자의 환기기능검사성적의 예측치. 29:510-4, 1986
- 조수현 등:연탄가스중독 발생실태에 관한 조사연구. 29:1233-40, 1986
- 최성기 등:쥐의 신동맥에서 단단문합술과 Telescoped 문합술의 비교에 관한 실험적 연구. 26:246-49, 1983
- 황성오 등:Raynaud 현상에서 수지온도 회복시간 측정에 관한 연구. 32:391-6, 1989

Table 1. Items of Checklist for Assessing the Methodological and Statistical Validity

연구방법론의 타당성 평가 항목:

1. 연구가설 또는 특수목표의 명백한 기술 (clear statement of research hypothesis or specific aims)
2. 적절한 초점 (suitable focus)
3. 모집단의 정의 (definition of population)
4. 적격성 기준의 명시 (eligibility criteria)
5. 제외 기준의 명시 (exclusion criteria)
6. 대표성 (representativeness)
7. 측정 방법의 상세한 기술 (description of methods in detail)
8. 타당도의 기술 (description of accuracy)
9. 신뢰도의 기술 (description of reliability)
- a10. 대조군의 선정 여부 (presence of control)
- a11. 감수성 비폴림 (susceptability bias)
- a12. 시행 비폴림 (performance bias)
- a13. 발견 비폴림 (detection bias)
- a14. 탈락 비폴림 (transfer bias)

통계처리기법의 타당성 평가 항목:

- *1. 기초자료의 불충분한 기술 (incomplete description of basic data)
- *2. 사용한 통계처리기법에 대한 불충분한 기술 (statistical test performed yet not defined)
- *3. 검정력, 신뢰구간에 대한 불충분한 기술 (incomplete description of power or confidence interval)
- **4. 대표치 및 산포도에 대한 부적절한 기술 (inadequate description of measures of central tendency or dispersion)
- ***5. 온당하지 않은 분석방법 (incorrect analysis)
- **6. 중복시행한 가설검정 (multiplicity on hypothesis testing)
- ***7. 부당하게 유도된 결론 (unwarranted conclusion)

Note) a; applicable for analytic study only

*; errors of omission

**; errors of commission