대 한 소 아 신 경 학 회 지
Vol. 11, No. 2, November, 2003

□원 저□

대한소아신경학회지 10년간(1993-2003)의 통계적 방법론에 대한 분석

계명대학교 의과대학 소아과학교실, 예방의학교실*

김준식 · 권태찬 · 이충원*

= **Abstract** =

An Assessment of Statistical Methods in the Journal of Korean Child Neurology Society, 1993-2003

Joon-Sik Kim, M.D., Tae-Chan Kwon, M.D. and Choong-Won Lee, M.D.*

Department of Pediatrics, Department of Preventive Medicine, Keimyung University, School of Medicine, Daegu, Korea*

Purpose : A clinical trial cannot be adequately interpreted without information about the methods used in the design of the study and the analysis of the results. We would like to evaluate trends in statistical methods and describe the frequency with which various statistical techniques are reported in the Journal of Korean Child Neurology Society.

Methods : We reviewed 288 original articles published in the Journal of Korean Child Neurology Society from 1993 to 2003 to assess the statistical methods.

Results : The number of cross-sectional study was 232(80.6%) articles and that of animal study was 45(15.6%) articles but Cohort study was only eleven(3.8%) articles. One hundred twenty seven(44.1%) articles used no statistical methods or descriptive statistics only, frequency of which decreased yearly and 84 articles(29.2%) used T-test, frequency of which increased yearly. Seventy two(25%) articles used contigency tables and twenty three(8.0%) articles used ANOVA. Orphan P where no statistical methods had been specified and only the P value given was presented in 19 articles(6.6%) which decreased yearly.

Conclusion : These results suggest that medical articles published in the Journal of Korean Child Neurology Society, 1993-2003, were short of their expected quality and the validity of statistical methods used appears to be seriously compromised in this period and has much to be done to improved the current situation. It is concluded that a basic training in biostatistical methods, more consultation of medical investigators with statistician or other experts, careful review by someone in biostatistics or research design before accepting a manuscript are needed.

Key Words : Journal of Korean Child Neurology Society, Biostatistical Methods

서 론

오늘날 의학분야에서 전문잡지를 통한 정보의 교환은 엄청나며, 이러한 연구는 질병의 새로운 기전이나 원인을 추론하거나, 어떤 약제나 치료법의 우월성을 증명하게 되는 데, 의학자나 임상의사들은 그 결과를 실제에 적용하기에 앞서 반드시 연구 논문의 질적 수준, 즉 방법론적 타당성 여부를 평가하여야 하게끔 되었다[1].

책임저자 : 김준식, 계명의대 동산의료원 소아과
Tel : 053)250-7525, Fax : 053)250-7783
E-mail : jskim@dsmc.or.kr

구미에서는 이에 대한 연구가 체계적으로 진행되어 연구방법론 및 통계 처리 기법의 타당성 여부를 검토하여 약 절반 정도에서 오류가 있었음을 밝혀내었으며 비록 평가 기준의 차이는 있지만 심지어 조사 대상 문헌의 5%만 타당한 것으로 보고한 예도 있다[2]. 국내에서는 의학 분야에서 몇 편의 통계학적인 타당도에 대한 연구들이 있으며[3, 4], 이 등[5]은 10년간 대한의학협회지에 발표한 논문 382편을 대상으로 타당성을 평가했을 때 연구의 타당성 점수가 60 이상인 원저는 14.9%에 지나지 않았고, 30 이하인 원저도 15.5%나 되었음을 지적하면서 연구방법론 및 통계처리기법상의 타당도가 아직 낮은 수준에 있는 것으로 평가한 바 있다.

그러나 1993년 창간된 대한소아신경학회지는 2003년 제11권 제1호까지 총 451편에 이르는 논문들이 발표되었으나, 지금까지 통계학적인 방법론에 대한 고찰이 시도된 적이 없었다. 이 연구는 1993년에서 2003년까지 10년간에 간행된 대한소아신경학회지에 수록된 논문을 대상으로 통계적인 방법론의 타당성을 분석 고찰함으로써 앞으로 대한소아신경학회지의 통계학적인 방법론의 질적 향상을 위한 기초 자료를 제공하고자 하며 더욱 논리적이고 과학적인 논문이 발표되는 학회지가 되기를 기대하면서 본 연구를 시행하였다

대상 및 방법

1993년에서 2003년 사이에 간행된 대한소아신경학회지에 게재된 논문 중 종설, 증례보고 및 기타 보고를 제외한 원저 288편을 대상으로 통계학적인 방법론의 측면에서 의학논문의 타당성을 평가하였다.

각 연구 논문의 방법론적 구조에 대한 분류를 하였으며 연구의 일반적 목적에 따라 분석적 연구와 기술적 연구, 시간축의 배열에 따라 경시적 연구(longitudinal study) 또는 코호트 연구와 단면연구(cross sectional study)로 나누었다. 논문의 구조적 분류와 함께 자료 분석에 이용한 통계적 기법에 대한 분류를 시행하여 백분율이나 히스토그램 외에 어떠한 통계적 내용도 담고 있지 않는 기술적 통계(descriptive statistics)만 사용한 경우, t 검정을 사용한 경우, X^2, Fisher exact 검사법 등 분할표를 사용한 경우, 상관계수를 사용한 경우, 회귀 분석을 사용한 경우, 분산 분석을 사용한 경우, 다중 비교를 분석한 경우, 기타의 통계를 이용한 경우와 사용한 통계처리를 밝히지 않은 경우로 나누어서 이들을 2년간 5개의 기간으로 나누어서 기간별로 증감을 보았다.

결 과

연도별 논문 수는 1999-2000년 사이에 격감한 것은 의약분업 파동에 따른 논문 수의 감소에 의한 것이며 2001-2003년 사이에는 상대적으로 5권의 대한소아신경학회지가 포함되어 증가된 것처럼 보이나 전반적으로 기간의 차이에 따른 논문 수의 증감은 없었다(Table 1).

연구방법은 횡단면적 연구가 232편(80.6%)으로 가장 많았으며 1993-1994년에는 91%(55편)이었으나 기간이 지나면서 70.8%(51편)로 감소하는 추세를 보였다. 동물 실험은 45편으로 15.6%이었으며 기간이 지나면서 늘어나는 경향을 보여 1993-1994년에는 3편(5%)에 불과하였으나 2001-2003년에는 20편(27.8%)으로 크게 증가하였다. 하지만 코호트 연구는 전체 11편으로 3.8%에 불과하였고 기간이 지나도 변화가 없었다(Table 2).

논문에 사용된 통계학적인 방법 중 가장 많은 빈도를 보인 것은 통계적인 방법을 사용하지 않았거나 또는 기술 통계(descriptive statistics)가 127편(44%)이었다. 다음은 t-검정이 77편(47.8%)이었으며 이중 독립 이표본 이분산 검정(independent

Table 1. Number of Articles Reviewed by Year

Years	Numbers(N=288)	Percent
93-94	60	20.1
95-96	65	22.6
97-98	54	23.7
99-00	37	12.8
01-03	72	25.0

Table 2. Frequency of Study Designs by Year

Methods	93-94 n=60(%)	95-96 n=65(%)	97-98 n=54(%)	99-00 n=37(%)	01-03 n=72(%)	Total n=288(%)
Cross-sectional	55(91.7)	57(87.7)	40(74.0)	29(78.4)	51(70.8)	232(80.6)
Cohort	2(3.3)	3(4.6)	3(5.6)	2(5.4)	1(1.4)	11(3.8)
Experiment	3(5.0)	5(7.7)	11(20.4)	6(16.2)	20(27.8)	45(15.6)

Table 3. Frequency of Use of Statistical Methods by Year

Methods	93-94 n=60(%)	95-96 n=65(%)	97-98 n=54(%)	99-00 n=37(%)	01-03 n=72(%)	Total n=288(%)
descriptive	32(53.3)	33(50.8)	21(38.9)	14(37.8)	27(37.5)	127(44.1)
T-test	11(18.3)	16(24.6)	10(18.5)	12(32.4)	35(48.6)	84(29.2)
Indepentdent	8(13.3)	10(15.3)	5(9.3)	5(13.5)	20(27.8)	48(16.7)
paired	0(0.0)	2(3.1)	1(1.9)	1(2.7)	5(6.9)	9(3.1)
Wilcoxon	2(3.3)	2(3.1)	1(1.9)	1(2.7)	6(8.3)	12(4.2)
Mann-Whitney	1(1.7)	2(3.1)	3(5.6)	5(13.5)	4(5.6)	15(5.2)
Contigency table	17(28.3)	14(21.5)	12(22.2)	13(35.1)	17(23.6)	72(25.0)
Chi-square	12(20.0)	10(15.3)	7(13.0)	9(24.3)	15(20.8)	53(18.4)
Fisher's exact	4(6.7)	4(6.2)	5(9.3)	4(10.8)	2(2.8)	19(6.6)
Mentel-Haezel	1(1.7)					
ANOVA	3(5.0)	3(4.6)	5(9.3)	4(10.8)	8(11.1)	23(8.0)
Correlation	6(10.0)	3(4.6)	3(5.6)		5(6.9)	16(5.6)
Pearson	2(3.3)	1(1.5)	1(1.9)	0	3(4.2)	7(2.4)
regression	3(5.0)	2(3.1)	2(3.7)	0	2(2.8)	9(3.1)
logistic	1(1.7)					
Survival	0	0	1(1.9)	1(2.7)	0	2(0.1)
unknown	2(3.3)	8(12.3)	8(14.8)	0	1(1.4)	19(6.6)

two sample t-test) 46편(16.7%), 짝진 이분산 검정(paired t-test)은 9편(3.1%)이었으며 비모수적인 방법인 Wilcoxon rank sum 방법이 12편(4.2%), Mann-Whitney 방법이 15편(5.2%)이었다. 다음으로 독립성 검정이 72편으로 25%를 차지하였으며 이중 Chi-square(X^2) 검정은 53편(18.4%), Fisher exact법이 19편(6.6%)이었다. T-검정과 독립성 검정에서 모수적(parametric) 방법과 비모수적(non-parametric) 방법으로 구분하였을 때 모수적 방법이 107편(72%), 비모수적 방법이 42편(28%)이었다. 분산 분석(ANOVA)은 23편(8%)이었으며 상관분석 및 회귀 분석은 15편(9%)이었고 생존 분석은 2편(1%)에 불과하였다(Table 3).

결과 또는 표에 P값만 제시해두고 어떤 통계기법을 사용하였는지 알 수 없는 경우가 19편으로 6.6%나 되었으며, t-검정이나 분산 분석의 구체적 방법 및 사후 검정에 대한 언급이 없어 통계적 기술로 부적절하였던 경우까지 포함하면 46편으로 전체 논문의 29%에 달하였다. 연도별로 보면 통계적인 방법을 사용하지 않았거나 기술 통계만을 사용한 것이 1993-1994년에 53.3%이었던 것이 매 기간마다 감소하여 2001-2003년에는 37%까지 감소하였으며, 이에 반하여 t 검정은 1993-1994년에 18.3%이었던 것이 매 기간마다 증가하여 2001-2003년에는 37%까지 증가하였으며, 결과 또는 표에 P값만 제시해두고 어떤 통계기법을 사용하였는지 알 수 없는 경우는 10%대에서 급격히 감소하였고, 분할표나 ANOVA는 기간별로 차이가 없었다.

고 찰

일반적으로 연구 논문의 가치는 실용적 가치와 학술적 가치로 나눌 수 있다[6]. 실용적 가치란 연구 결과를 실제로 적용했을 때 기대되는 효과의 크기를 말한다. 학술적 가치는 다시 창의성(연구 주제가 얼마나 참신하고 풍부한 상상력에 근거한 것인가?), 기술적 난이도(연구에 사용된 측정 방법이나 처치 등이 얼마나 고도의 기술을 필요로 하는 것인가?), 연구방법론의 타당성으로 분류할 수 있다. 자신의 연구 결과나 주장이 타당하다는 것을 제시하기 위해서 먼저 충족되어야 하는 전제조건을 연구의 계획에서부터 수행, 결과분석 및 결론 도출에 이르는 모든 연구 과정에 객관적이고 과학적인 방법을 적용하는 것이며[4], 특히 의학적인 연구에서 통계기법을 포함한 연구방법론의 타당성이 의심될 때, 논문의 결과와 이에 대한 저자의 오류를 범한 해석은 연구방법론과 통계적인 방법에 무지한 독자들에게 잘못된 정보(misinformation)를 전달하게 되어 결과적으로는 환자 관리(patient management)에 심각한 문제의 소지를 남기게 된다. 그래서 연구의 목적 및 자료의 성상에 적절한 연구방법과 통계적인 적용 및 해석은 의학적인 논문에서 점점 필수적인 요건이 되어 가고 있다.

연구방법론에서 전 세계적인 추세가 단순한 임상적 고찰이나 횡단면적인 연구에서 벗어나 분석적인 연구, 특히 코호트 연구와 임상시험으로 이행된 단계이므로 이를 다룬 임상역학(clinical epidemiology)에 대한 개념 확산이 되어 원인 연구(etiology study)에 주력해야 함이 시급함을 알 수 있다. 임상적인 고찰을 하는 이유는 보기 드문 질병에 대해서 기술통계를 보아 가설을 설정하고(hypothesis generation) 원인 연구를 시행할 전단계의 연구(precursor)로서 의미를 지니고 있어야 한다[7]. 재료 및 방법에는 반드시 연구방법론을 구분해서 기술을 해야 하며 연구결과를 해석할 때는 사용한 연구방법론에 적절하게 해야 한다. 인과관계를 정립할 수 없는 연구방법론을 사용하고서 인과관계를 보이는 것처럼 연구결과를 확대 해석해서는 곤란할 것이다. 그리고 동물의 실험연구에서는, 실험방법론에 입각한 실험과 통계적인 분석을 하게 되면 적은 비용으로 최대한 가설검정을 할 수 있어 경제적인 연구가 될 수 있으므로[8] 이에 대한 적절한 이해가 요구된다 .

이 등[5]이 1980년 1월부터 1989년 12월까지 대한의학협회지에 발행된 원저 382편의 분석에서 기술적 연구, 조사연구, 단면연구에 비해 과학적으로 더욱 정밀성이 요구되는 분석적 연구, 실험, 경시적인 연구가 전체적으로 40.6%, 12.8%, 17.3%에 불과하였으며, 10년 추이에서도 연구방법론의 변화 양상이 뚜렷하지 않았다고 보고한 바 있다. 그러나 Feinstein[9]은 1977년과 1978년 Lancet와 New England Journal of Medicine에 발표된 311편의 원저에서 분석적 연구(60%), 실험(20%), 경시적인 연구(39%)였다고 보고한 바 있다. 연구방법론에 대한 정확한 이해가 선행되어야 하고 논문에 이를 정확하게 기술하고 이에 따른 해석을 해야 하나 DerSimonian 등[10]은 New England Journal of Medicine, Lancet, British Medical Journal, JAMA의 1979년에 발행된 67편의 임상시험을 평가하였을 때, 80%에서 통계적인 분석, 사용된 통계적인 기법, 확률할당에 대해 밝히고 있으나 19%만이 확률할당의 방법에 대해 기술했고 추적손실은 79%, 치료 순응도는 64%, 등록기준(eligibillity criteria)은 37%, 통계적인 검정력은 12%에서 기술했다고 보고하였다.

대한소아신경학회지에서는 분석적인 연구방법에 속하는 코호트 연구는 전체 11편으로 3.8%에 불과하였고 기간이 지나도 변화가 없었으나 동물 실험은 15.6%이었으며 기간이 지나면서 늘어나는 경향을 보여 1993-1994년에는 5%에 불과하였으나 2001-2003년에는 27.8%로 크게 증가하였다.

이 등[5]은 분석한 총 382편 중 통계적인 처리방법을 밝히지 않은 논문이 117편으로서 통계적인 방법을 사용한 221편을 분모로 하면 52.9%나 되며, 통계처리 기법을 밝히지 않은 경우를 제외하면 t 검정이 11.8%, Pearson의 상관분석이 11.0%, 분

할표(contingency table, X^2 검정)가 9.4%로 대부분 간단한 단일분석이 대다수를 차지하고 있음을 보고하였다. 국외의 예로는 Felson 등[11]이 1967-1968년과 1982년에 Arthritis and Rheumatism에 게재된 논문의 연구방법을 비교해 보았을 때, 통계적인 방법을 사용한 논문이 50%(47/94)에서 62%(74/119)로 증가하였으며, t 검정과 X^2 검정을 사용한 논문의 비율이 각각 17%에서 50%, 19%에서 30%로 증가했음을 보고하였다. 또 선형회귀분석(linear regression)은 2%(1편)에서 24%로 증가했으며 하나 이상의 통계적인 기법을 사용한 논문의 비율은 9%에서 41%로 증가하였다고 하였다. Altman[12]이 1990년 New England Journal of Medicine지에 발행된 순서대로 평가한 100편의 논문을 동일한 잡지의 1978-1979년에 게재된 논문을 분석한 Emerson과 Colditz[13]의 결과와 비교를 했을 때, 기술통계만을 사용한 논문이 27%에서 11%로 감소했으나 간단한 단일변수기법은 거의 변화가 없었다고 보고하였다. 그러나 선형회귀분석과 비모수 검정법(non-parametric methods)은 거의 2배의 증가를 보였으며, 좀더 복잡한 다변수 분석은 극적인 증가를 보였다. 큰 폭으로 증가한 기법은 생존분석법(survival analysis)으로서 100편 중 27편에서 이 방법을 사용했는데 대부분 다중로지스틱회귀분석(multiple logistic regression)과 Cox 회기분석법(proportinal hazard method)의 증가였다고 보고하였다. 논문 편수당 서로 다른 통계적인 기법을 사용한 횟수가 증가하는 경향을 보였다고 하였다. 분산분석의 사용은 비교적 제한된 상황인데, 암 관련 원저에서 2% 정도, 정신과 잡지에서는 거의 10%였다[14].

대한소아신경학회지에서는 통계적인 방법을 사용하지 않았거나 또는 기술 통계(descriptive statistics)가 44%이었으나 매 기간마다 점차적으로 감소하여 1993-1994년에 53.3%이었던 것이 매 기간마다 감소하여 2001-2003년에는 37%가 되었다. t 검정은 1993-1994년에 18.3%이었던 것이 매 기간마다 증가하여 2001-2003년에는 37%까지 증가하여 Felson 등[11]과 유사하였으며, 분산 분석은 8%로 Hokanson 등[14]의 보고와 유사하였다. 그리고 결과 또는 표에 P값만 제시해두고 어떤 통계기법을 사용하였는지 알 수 없는 경우가 19편으로 6.6%나 되었으며, t-검정이나 분산 분석의 구체적 방법 및 사후 검정에 대한 언급이 없어 통계적 기술로 부적절하였던 경우까지 포함하면 46편으로 전체 논문의 29%에 달하였다.

자료분석의 전 세계적인 추세가 단일분석(univariate analysis), 층화분석(stratified analysis)의 단계를 지나 현재는 다변수분석의 단계에 접어들었으므로[15] 이의 적용이 시급함을 알 수 있다. 특히 여러 변수에 대해서 두 그룹 간에 단일 변수로 유의성 검정을 하게 되면 변수들 간의 상관성을 고려하지 못하게 되므로 통계적인 제1종 오차의 가능성이 커지므로 다변수 분석을 권고하고 있는 상황이다. 그러나 다변수 분석은 그 자체가 까다로운 기본적인 가정과 통계학적인 모델링(statistical modelling)의 문제를 가지고 있으므로 적용시 주의를 요한다[15].

최 등[16]은 최근에 통계적인 기법들이 다양해지고 내용 또한 그 수준이 높아지고 있다고 평가하면서, 1983-1987년 사이에 교육학연구와 한국영양학회지에 수록된 논문 중 점검표를 이용해서 통계적인 기법을 활용한 논문 총 35편에 대한 타당도를 조사했을 때 논문 모두가 정도의 차이가 있을 뿐 통계적인 기법 활용에 문제점을 지니고 있다고 보고하였다. 단계별로 제시한 문제점들을 보면, 먼저 연구 설계 과정에서는 연구 대상에 대한 사전 탐사의 부족, 대표성의 고려 및 표본 크기에 대한 언급의 부족, 자료탐사의 무시, 실험 설계의 미숙함, 측정의 문제 등이 있었으며, 통계적인 추론의 단계에서는 적절한 통계적 기법 선택의 문제와 적용 절차상의 문제, 유의 수준의 적용문제, 활용한 통계패키지의 구체적인 언급의 부족, 연구 과제의 통계적인 형식화의 고려부족 등이었다. 또한 마지막 단계인 결론 도출 과정에서는 연구가설에 대한 결론의 기술문제가 가장 문제가 되었는데 결론의 서술시 이를 정당화시키는데 필요한 통계량이나 P값이 제시되고, 이를 바탕으로 엄격한 서술이 이루

어져야 하나, 대부분의 논문이 이 점을 소홀히 했으며, 특히 결론 기술을 너무 단정적으로 표현하는 것도 문제점으로 지적한 바 있다.

이 연구에서 통계적인 기법의 적용이 잘못된 대표적인 예로 반복측정을 했을 때 처음 측정한 기초측정치(baseline measure)를 대조군으로 두고 각 시간별 측정치를 짝 비교 t 검정(paired t-test)을 반복해서 실시한 경우와 3 그룹 이상의 평균치 비교에 독립 t 검정(independent t-test)시 문제는 통계적인 검정시에 항상 범하게 되는 제 1종 오차(type I error)가 커져서 실제로는 유의하지 않은 그룹간의 유의한 것으로 나타날 가능성이 커진다는 것이며(error inflation), 이를 방지하기 위해서는 분산 분석을 실시한 후 연구목적에 적합한 반복측정 분석(repeated measurement)이나 공변량 분석(ANCOVA)과 같은 다중비교(multiple comparisons)를 하는 것이 권장할 방법이다[17]. 둘째로 모수적인 통계기법을 사용한 상당수의 논문이 비모수적인 기법을 사용하는 것이 더 적절했을 것으로 판단되었다. 비모수검정은 보통 자료가 명백하게 정규분포를 취하지 않을 때, 표본크기가 너무 작아서 자료의 분포를 알 수 없을 때, 빠르게 결과를 알고 싶을 때 그리고 자료가 순서척도일 때 사용하는 방법으로서 보통 모수검정법보다 검정력이 떨어지는 방법으로 알려져 있다. 그러나 위와 같이 모수적인 방법을 적용시킬 수 없을 때 적용시켜서 얻을수 있는 검정력보다는 비모수검정법을 적용시키는 것이 더 강력한 검정력을 얻을 수 있고, 특히 짝비교 t 검정에 대한 비모수검정법인 Wilcoxon signed ranks test는 정규분포를 취할 때는 짝비교 t 검정과 동등한 검정력을 가지고, 만약에 비정규분포를 취할 경우에는 보다 더 강력한 검정력을 가지는 통계기법으로 의학잡지에서 점차적으로 많이 사용되고 있는 기법이므로 비모수검정법 사용에 대한 인식이 달라져야 할 것이다[18]. 셋째, 역학적인 통계기법의 하나인 비교위험도(relative risk) 또는 대응비(odds ratio)를 사용한 논문은 거의 없었으며, 혼란변수에 대한 개념과 통제의 방법에 대한 기술 역시 거의 없어서 임상역학에 대한 개념정립이 시급함을 보여주었다. 재료 및 방법에서는 최소한 사용한 통계적인 기법, 사용에 따른 기본적인 가정을 충족되는지의 여러 통계적인 유의성을 결정하는 유의수준 등에 대한 기술이 되어야 한다.

통계기법의 기본적인 가정에 어긋남에도 불구하고 적합하지 않은 통계기법을 사용하게 되면 잘못된 결과와 이에 따른 해석이 도출되어 논문이 결론에 심각한 문제를 야기시킬 수 있으므로 어떠한 통계기법이라도 반드시 이에 대한 점검이 있은 후에 적용을 시켜야 할 것이다. 통계적인 유의성을 선언하는 기준이 되는 유의수준(significance level)의 기술 역시 총 12편으로 9.8%에 지나지 않아 문제가 되었다. 통계적인 유의수준은 표본의 크기와 표준편차에 따라 좌우되므로 이에 대한 기술 역시 있어야 한다. 특히 연구시작 전에 계획중인 연구가 통계적인 유의성을 가지기 위해 필요한 표본크기를 사전에 계산해 보고 이에 맞추어 대상자를 얻어야 하나[19] 이러한 논문은 거의 없었다. 또 연구 결과가 기대했던 결과를 나타내지 못 했을 때(negative findings) 시행된 연구의 표본크기에 따른 통계적인 검정력(power)에 대한 점검을 해보아 통계적인 제 2종 오차(type II error)의 정도를 고려해 보아야 하나[20], 이를 고찰한 논문 역시 거의 없었다. 현재는 개인용 컴퓨터를 이용해서 대부분의 연구방법과 통계기법에 대해 표본크기와 검정력을 계산해주는 프로그램이 있으므로 연구를 시작하기 전에 반드시 표본크기를 계산한 후 대상자 선정과 등록의 단계에 들어가야 할 것이다. 이는 특히 임상시험(clinical trial)일 경우 중요한 의미를 지닌다. 그리고 통계적인 유의성(statistically significance)이 항상 임상적인 유의성(clinical significance)을 의미하는 것이 아니므로 통계적인 유의성이 연구결과를 해석하는 데 있어 절대적인 기준이 되지 못함을 알아야 한다. 특히 점추정(point estimation)으로서의 P값은 큰 의미를 지니지 못하나[21] 최근 들어 통계적인 기법이 우리나라 의학계에 도입이 되면서 연구자들이 너무 P값에 의존하는 경향을 보이고 있다. 신뢰구간은 독자들에게 연구에서 제시된 추정치가 변이를 가지고 있어 만

약에 연구를 재현(replication)한다면 동일한 결과를 얻을 수 없을 것이라는 것을 상기시켜 줄 수 있다는 점과 가설검정이 제시해주는 정보 이상의 것을 제공해준다는 점에서 많이 이용되고 있다[18].

국외의 저명한 의학잡지에서는 통계적인 검정시 *P*값에 너무 의존하는 경향에서 벗어나 신뢰구간(confidence intervals)의 추정으로 방향을 전환하고 있는 추세이다. British Medical Journal의 편집자들은 신뢰구간이 적절한 경우에는 가설검정대신에 신뢰구간을 제시하도록 하는 방침을 정해 두고 있다[22].

이러한 통계적인 적용상의 문제점들을 해결하여 타당도를 향상시키기 위해서 첫째, 대학과 대학원 과정에서 올바른 통계학적인 교육이 선행되어야 하고 둘째, 게재를 위해 제출된 논문에 대해서 전문가들이 사독과정(review process)을 거쳐서 조언을 해줄 수 있는 체계(referee system)를 확립하고 셋째, 통계적인 타당도는 연구의 디자인과 불가분의 관계를 가지고 있으므로 연구를 디자인하는 단계에서부터 통계 전문가들과 협력해서 연구 종료시에 타당한 통계적인 적용과 해석을 할 수 있도록 도와줄 수 있는 통계상담(statistical counselling) 서비스 체제의 완비와 이를 활용하고자 하는 연구자들의 의지가 필요하고 넷째, 논문집에 독자의 반론을 실을 수 있는 지면의 마련 등이 제시되고 있다[16]. 투고규정에 최소한 형식적으로 갖추어야 할 통계학적인 기준을 제시하고 이를 어겼을 경우에는 다시 수정하도록 하는 제도 역시 고려해 볼만하다.

이 연구의 한계점으로는 연구의 타당성 조사가 제대로 이루어지지 않았고, 연구 설계과정, 표집과정, 통계적인 추론, 결론 도출과정에 대해 자세하게 세분해서 평가를 하지 않아 현실적인 문제점으로 객관적이고 형식적인 부분에 대해서만 평가를 했다는 점이며 이에 대한 연구가 더 필요할 것으로 사료된다.

요 약

목 적 : 1993년에서 2003년 사이에 간행된 대한소아신경학회지에 게재된 논문에 사용된 연구 방법과 통계적 방법론의 변화와 오류를 분석하여 더 나은 학회지가 되도록 연구를 시행하였다.

방 법 : 1993년에서 2003년 사이에 간행된 대한소아신경학회지에 게재된 논문 중 원저 288편을 대상으로 통계학적인 방법론을 분석하였다.

결 과 : 연구 방법은 횡단면적 연구가 232편(80.6%)으로 가장 많았으며 1993-1994년에는 91%(55편)이었으나 기간이 지나면서 70.8%(51편)으로 감소하는 추세를 보였다. 동물 실험은 45편으로 15.6%이었으며 기간이 지나면서 늘어나는 경향을 보여 1993-1994년에는 3편(5%)에 불과하였으나 2001-2003년에는 20편(27.8)으로 크게 증가하였다. 하지만 코호트 연구는 전체 11편으로 3.8%에 불과하였고 기간이 지나도 변화가 없었다. 논문에 사용된 통계학적인 방법 중 가장 많은 빈도를 보인 것은 통계적인 방법을 사용하지 않았거나 또는 기술 통계(descriptive statistics)가 127편(44%)이었다. 다음은 t-검정이 77편(47.8%)이었으며 독립성 검정이 72편으로 25%이었으며 분산 분석(ANOVA)은 23편(8%)이었으며 상관 분석 및 회귀 분석 15편(9%)이었고 생존 분석 2편(1%)에 불과하였다. 결과 또는 표에 *P*값만 제시해두고 어떤 통계기법을 사용하였는지 알 수 없는 경우가 19편으로 6.6%나 되었으며, t-검정이나 분산 분석의 구체적 방법 및 사후 검정에 대한 언급이 없어 통계적 기술로 부적절하였던 경우까지 포함하면 46편으로 전체 논문의 29%에 달하였다. 연도별로 보면 통계적인 방법을 사용하지 않았거나 기술 통계만을 사용한 것이 1993-1994년에 53.3%이었던 것이 매 기간마다 감소하여 2001-2003년에는 37%까지 감소하였으며, 이에 반하여 t 검정은 1993-1994년에 18.3%이었던 것이 매 기간마다 증가하여 2001-2003년에는 37%까지 증가하였으며, 결과 또는 표에 *P*값만 제시해두고 어떤 통계기법을 사용하였는지 알 수 없는 경우는 10%대에서 급격히 감소하였다. 상대적으로 횡단면적 연구와 기술 통계에 그친 논문이 많았으며, 통계학적 방법을 이용하더라도 간단한 t-검정이나 독립성 검정이 주류를 이루었고, 다중비교는

거의 볼 수가 없었다.

결 론: 대한소아신경학회지의 게재될 논문의 질을 높이기 위해서 연구에 적절하고 합리적인 통계방법이 적용되어야 하겠고 통계적인 타당성의 향상을 위해서는 논문 게재를 원하는 저자와 편집자 양측 모두가 상당한 노력이 필요할 것임을 시사하였다.

참 고 문 헌

1) Lionel ND, Herxheimer A. Assessing reports of therapeutic trials. Br Med J 1970;3:637-40.
2) Mahon WA, Daniel EE. A method for the assessment of reports of drug trials. Can Med Assc J 1964;90:565-9.
3) 안윤옥, 고응린. 자료처리과정에 대한 통계학적 검토-일부 의학잡지에 게재된 논문예를 중심으로. 예방의학회지 1973;6:81-5.
4) 안윤옥, 이형기. 의학에서의 연구방법론. 한국역학회지 1990;12:107-14.
5) 이형기, 허봉열, 안윤옥. 1980년대에 발표된 국내 의학연구 논문의 방법론 및 통계처리기법의 타당성에 관한 평가연구. 가정의 1991;12:46-67.
6) 김해동. 조사방법론. 재판. 서울 : 법문사, 1988: 69-137.
7) Fletcher RH, Fletcher SW, Wagner EH. Clinical epidemiology : The essentials, ed 2. Baltimore, Williams & Wilkins, 1988:188-207.
8) Kelinger FN. Foundations of behavioral research, 3rd ed. New York : Holt, Rinehart and Winston, 1986:279-346.
9) Feinsten AR. Clinical biostatistics : A survey of the research architecture used for publications in general medical journals. Clin Pharmacol Ther 1978;23:117-25.
10) DerSimonian R, Charette LJ, McPeek B, Mosteller Fl. Reporting on methods in clinical trials. N Engl J Med 1982;306:1322-7.
11) Felson DT, Cupples LA, Meenan RF. Misuse of statistical methods in arthritis and rheumatism. Arthritis Rheum 1984;27:1018-22.
12) Altman DG. Statistics in medical journals : Developments in the 1980s. Stat Med 1991;10: 1897-913.
13) Emerson JD, Colditz GA. Use of statistical analysis in the New England Journal of Medicine. N Engl J Med 1983;309:709-13.
14) Hokanson JA, Luttman DJ, Weiss GB. Frequency and diversity of use of statistical techniques in oncology journals. Cancer Treat Rep 1986;70:589-94.
15) Hanley JA. Appropriate uses of multivariate analysis. Ann Rev Public Health 1983;4:155-80.
16) 최종후, 김기목, 김기영. 과학 학술지에 나타난 통계적 기법활용의 타당성 평가. 응용통계 1990; 5:1-13.
17) Smith DG, Clemens J, Crede W, Harvey M, Gracely EJ. Impact of multiple comparisons I randomized clinical trials. Am J Med 1987;83: 545-50.
18) Dawson-Saunders B, Trapp RG. Basic and clinical biostatistics. East Norwalk, Prentice-Hall International Inc, 1990:64-98, 110-1, 207-28, 264-75.
19) Kraemer HC. Sample size : When is enough? Am J Med Sci 1988;296:360-3.
20) Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of theta, the type II error and sample size in the design and interpretation of the randomized control trial : Survey of 71 "Negative" trials. N Engl J Med 1978;299:690-4.
21) Diamond GA, Forrester JS. Clinical trials and statistical verdicts : Probable grounds for appeal. Ann Int Med 1983;98:385-94.
22) Gardner JM, Altman DG. Confidence intervals rather than P values : Estimation rather than hypothesis testing. Br Med J 1986;292:746-50.