啓明醫大論文集 第12卷 第4號
The Keimyung Univ Med J
Vol 12, No. 4, December, 1993

계명의대 논문집 10년간(1982－1991)의 통계적 방법론에 대한 고찰*

계명대학교 의과대학 예방의학교실

윤능기 · 이충원 · 서석권 · 박종원

서 론

의학분야에 통계적인 방법이 도입된 이후로 그 적용의 수가 최근들어 비약적으로 증가하였지만 그 양에 비해 질적인 문제는 논의의 대상이 되고 있으며 구미에서는 의학잡지에서 잘못 사용된 통계적인 기법에 대한 우려가 있은지 최소한 60년 이상이 되었다(Altman, 1991). 이에 따라 구미에서는 일찍부터 이 분야에 대해 체계적으로 연구가 진행되어 왔다(Dunn, 1929; Greenwood, 1932; Schor와 Karten, 1966; Schoolman 등, 1968; Gardner, 1975; Glantz, 1980; DerSimonian 등, 1982; Felson 등, 1984; Evans와 Pollock, 1985; Pocock, 1985; Thoron 등, 1985; Williamson 등, 1986; Evans와 Pollock, 1987; Chalmers, 1988; Hebert와 Miller, 1988; Andersen, 1990). Williamson 등(1986)은 발행된 의학연구논문의 연구방법, 자료수집, 통계적인 기법의 과학적인 적절함을 평가한 28편의 종설을 검토했을 때 평균적으로 4,235편의 논문 중 약 20%에서만 평가자들의 타당도 기준을 만족시킨 것으로 보고한 바 있다. 국내에서는 의학분야에서 몇 편의 통계학적인 타당도에 대한 연구들이 있다(안윤옥과 고응린, 1973; 근로복지공사중앙병원부설 직업병연구소, 1990; 안윤옥과 이형기, 1990; 최종후와 이재창, 1990; 최종후 등, 1990; 이형기와 안윤옥, 1991). 이형기와 안윤옥(1991)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 연구 방법론 및 통계처리기법상의 타당도가 아직 낮은 수준에 있는 것으로 평가한 바 있다. 그러나 1982년 창간된 계명의대논문집은 1991년 재 10권 제4호까지 총 438편에 이르는 논문들이 발표되었으나, 지금까지 통계학적인 방법론에 대한 고찰이 시도된 적이 없었다.

이 연구는 1982년에서 1991년 10년간에 간행된 계명의대 논문집에 수록된 논문을 대상으로 통계적인 방법론의 타당성을 분석고찰해서 앞으로 계명의대 논문집의 통계학적인 방법론의 질적향상을 위한 기초자료를 제공하고자 한다.

재료 및 방법

1982년에서 1991년 사이에 간행된 계명의대잡지를 대상으로 게재된 논문 438편에서 단순한 임상적 고찰(case-series), 형태학적인 연구, 종설, 증례보고 및 기타 보고를 제외한 원저 204편을 대상으로 통계학적인 방법론의 측면에서 의학논문의 타당성을 평가하였다. 타당도를 평가하는 기준은 기존에 개발되어 있는 국내외의 점검표(checklist)를 참고(최종후 등, 1990; 안윤옥과 이형기, 1991)하여 이 연구에 적합한 점검표를 완성한 후 기준에 따라 평가하였다. 점검표의 평가에는 계명대학교 의과대학 예방의학교실과 외부 통계학과의 대학원생이 참여하였다. 점검항목은 연구방법, 통계적인 방법, 표본추출방법, 재료 및 방법에 통계적인 방법, 기본적인 가정, 유의수준 등을 기술했는지 여부, 통계적인 적용의 타당성 등이었으며 이들을 2년간 5개의 기간으로 나누어서 기간별로 증감을 보았다. 연구방법은 임상적 고찰, 단면적 연구, 환자－대조군 연구, 코호트연구, 임상시험, 인간과 동물을 대상으로 한 실험으로 분류를 했으며, 표본추출방법은 대표성을 지닌 표본추출법, 목적적 또는 임의 표집법으로 분류를 하였으며, 동물실험 61편을 제외한 143편을 대상으로 하였다. 통계적인 방법의 기술 여부와 통계적인 방법 적용의 적절성을 볼 때는 통계적인 방법과 무관한 82편의 논문을 제외한 122편의 논문을 대상으로 하였다.

* 이 논문은 1993년도 동산의료원 특수과제 연구비로 이루어졌음.

결 과

점검한 총 논문수는 204편으로 2년 간격으로 연도별로 보았을 때 기간이 증가함에 따라 논문 수가 증가했으며, 특히 88-89년 53편, 90-91년에 55편으로 상당히 증가했음을 알 수 있었다(표 1).

연구방법은 횡단면적인 연구가 81편(39.7%)으로 가장 많았으며 동물실험이 61편(30.0%)이었다. 코호트 연구는 4편(1.9%)에 지나지 않았다. 연도별로 보았을 때 전기간에 걸쳐서 횡단면적인 연구가 가장 많았으며 82-83년에는 8편(38.1%), 84-85년에는 10편(32.3%)이다가 86-87년부터는 40% 이상이 되었다. 다음으로는 동물실험이 많았는데 82-83년에는 9편으로 42.9%를 차지하였으나 그 이후로 점차 감소하여 90-91년에는 27.3%를 나타내었다. 임상적인 고찰 역시 10% 내외의 비율을 보이다가 90-91년에는 1편이었다. 코호트 연구는 87년까지 없다가 88-89년에 1편, 90-91년에 3편이었으며 임상시험은 84년부터 나타났으나 이후에도 10% 미만의 편수를 보였다(표 2).

논문에 사용된 통계학적인 방법 중 가장 많은 빈도를 보인 것은 통계적인 방법을 사용하지 않았거나 또는 기술통계만을 사용한 것이 87건(36.1%)이었다. 다음은 t-test로서 68건(28.2%)이었다. 이중 짝비교는 15건이었다. 비모수검정과 생존분석 그리고 다변수분석인 다중회귀분석(multiple linear regression)은 각각 1건(0.4%)에 지나지 않았다. 결과 또는 표에 p 값만 제시해두고 어떤 통계기법을 사용했는지 알 수 없는 경우가 25건으로 10.4%나 되었다. 연도별로 보면 통계적인 방법을 사용하지 않았거나 또는 기술통계만을 사용한 것이 82-83년에 68.2%였던 것이 매 기간마다 조금씩 감소해서 90-91년에는 20.8%로 감소하였으나 반면에 사용한 통계기법을 모르는 경우는 기간별로 감소 추세를 보이지 않았다(표 3).

대표성을 지닌 표본 표집법을 사용한 논문은 1편에 지나지 않았고 임의 표집법 또는 목적적 표집법이 114편(79.7%)으로 대다수를 차지하였다. 그리고 표본 표집법에 대한 기술이 없거나 잘 모를 경우는 28편(19.6%)이었다. 연도별로는 임의 또는 목적적 표집법이 82-83에는 9편(75.0%)이었으나 84-85년에 17편(70.8%)으로 증가했으며, 86년 이후로는 95%를 넘었다. 반면에 기술되지 않았거나 또는 잘 모를 경우가 82-83년에는 3편(25.0%)이었고 84-85년에는 6편(25.0%)이었으나 86-89년부터는 3%내외로 급격하게 감소하였다(표 4).

재료 및 방법에 기본적인 통계학적인 방법에 대한 기술을 해 두었는지를 볼 때 분모는 통계학적인 방법과는 관련이 없는 82편을 제외한 122편이 되었다. 사용한 통계학적인 방법을 기술한 논문은 73편으로 59.8%에 지나지 않았으며, 통계적인 방법을 적용시킬 때 필요한 기본적인 가정을 기술하거나 사용한

Table 1. Number of articles reviewed by year

Years	Number(N=204)	Percent
82-83	21	10.3
84-85	31	15.2
86-87	44	21.6
88-89	53	26.0
90-91	55	27.0

Table 2. Frequency of study designs by year

Methods	82-83 (21)	84-85 (31)	86-87 (44)	88-89 (53)	90-91 (55)	Total (204)
Case-series	-	4(12.9)	4(9.1)	7(13.2)	1(1.8)	16(7.8)
Cross-sectional	8(38.1)	10(32.3)	19(43.2)	22(41.5)	22(40.0)	81(39.7)
Case-control	1(4.8)	5(16.1)	2(4.5)	3(5.7)	8(14.5)	19(9.3)
Cohort study	-	-	-	1(1.9)	3(5.5)	4(1.9)
Clinical trial	-	2(6.5)	3(6.8)	3(5.7)	5(9.1)	13(6.4)
Experiment(human)	3(14.3)	3(9.7)	2(4.5)	1(1.9)	1(1.8)	10(4.9)
Experiment(animal)	9(42.9)	7(22.6)	14(31.8)	16(30.2)	15(27.3)	61(30.0)

Table 3. Frequency of use of statistical methods by year

Methods	82-83 (21)	84-85 (31)	86-87 (48)	88-89 (68)	90-91 (72)	Total (241)
No statistical method or descriptive statistics	15(68.2)	18(58.1)	20(41.7)	19(27.9)	15(20.8)	87(36.1)
T-test						
Independent	-	7(22.6)	12(25.0)	14(20.6)	20(27.8)	53(22.0)
Paired	2(9.1)	1(3.2)	4(8.3)	3(4.4)	5(6.9)	15(6.2)
Chi-sqaure	-	3(9.7)	1(2.1)	3(4.4)	4(5.6)	11(4.6)
Pearson correlation	1(4.5)	1(3.2)	2(4.2)	4(5.9)	6(8.3)	14(5.8)
Non-parametric tests	-	-	-	-	1(1.4)	1(0.4)
Survival analysis	-	-	-	1(1.5)	-	1(0.4)
Epidemiologic statistics	-	-	1(2.1)	3(4.4)	2(2.8)	6(2.5)
Simple linear regression	1(4.5)	-	1(2.1)	3(4.4)	4(5.6)	9(3.7)
Analysis of variance	-	-	3(6.3)	8(11.8)	7(9.7)	18(7.5)
Multiple linear regression	-	-	-	-	1(1.4)	1(0.4)
Unknown	3(13.6)	1(3.2)	4(8.3)	10(14.7)	7(9.7)	25(10.4)

Table 4. Frequency of sampling methods by year

Methods	82-83 (12)	84-85 (24)	86-87 (30)	88-89 (37)	90-91 (40)	Total (143)
Representative	-	1(4.2)	-	-	-	1(0.7)
Accidental or purposive	9(75.0)	17(70.8)	29(96.7)	36(97.3)	39(97.5)	114(79.7)
Not described or unknown	3(25.0)	6(25.0)	1(3.3)	1(2.7)	1(2.5)	28(19.6)

이유를 제시한 논문은 단지 3편으로 2.5%에 지나지 않았다. 논문에서 저자가 적용한 통계적인 유의수준(significance level)을 기술한 논문은 12편(9.8%)에 지나지 않았다. 이를 연도별로 보면 통계적인 방법의 기술을 한 논문은 82-83년에 1편이었으나 84-85년에 5편으로 31.3%, 86-87년에는 14편으로 53.4%, 88-89년 이후에는 70%를 상회하였다. 기본적인 가정을 기술한 논문은 총 3편에 지나지 않아서 연도별 비교가 무의미하였다. 유의성 수준의 기술은 82-83년에는 2편으로 22.2%, 84-85년에는 5편으로 31.3%였으나 86-87년에 1편으로 3.8%를 나타내어 오히려 급격히 감소하는 경향을 보였다. 88-89년, 90-91년에도 각각 3.1%, 7.7%를 나타내어 별 차이를 보이지 않았다(표 5).

통계적인 기법의 적용이 연구의 목적과 디자인, 자료의 성상 등에 적절한 것이었는지를 보면 30편(24.6%)만이 적절했고 나머지 92편(75.4%)에서 부적절한 것으로 평가되었다. 연도별로는 적절한 논문이 82-83년의 1편에서 점차적으로 증가하여 88-89년에 13편(40.6%)으로 증가하였으나 90-91년에 다시 9편(23.1%)으로 감소하였다(표 6).

고 찰

자신의 연구결과나 주장이 타당하다는 것을 제시하기 위해서 먼저 충족되어야 하는 전제조건은 연구의 계획에서부터 수행, 결과분석 및 결론 도출에 이르는 모든 연구 과정에 객관적이고 과학적인 방법을 적용하는 것이다(안윤옥과 이형기, 1990). 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

Table 5. Descriptions of statistical methods in the articles by year

Methods	82-83 (9)	84-85 (16)	86-87 (26)	88-89 (32)	90-91 (39)	Total (122)
Statistical methods						
Described	1 (11.1)	5 (31.3)	14 (53.8)	23 (71.9)	30 (76.9)	73 (59.8)
Not described	8 (88.9)	11 (68.7)	12 (46.2)	9 (28.1)	9 (23.1)	49 (40.2)
Basic assumptions						
Described	1 (11.1)	-	-	1 (3.1)	1 (3.1)	3 (2.5)
Not described	8 (88.9)	16(100.0)	26(100.0)	31 (96.9)	38 (96.9)	119 (97.5)
Significance level						
Described	2 (22.2)	5 (31.3)	1 (3.8)	1 (3.1)	3 (7.7)	12 (9.8)
Not described	7 (77.8)	11 (68.7)	25 (96.2)	31 (96.9)	36 (92.3)	110(90.2)

Table 6. Appropriateness of application of statistical methods by year

Methods	82-83 (9)	84-85 (16)	86-87 (26)	88-89 (32)	90-91 (39)	Total (122)
Proper	1(11.1)	1(6.3)	6(23.1)	13(40.6)	9(23.1)	30(24.6)
Not proper	8(88.9)	15(93.7)	20(76.9)	19(59.4)	30(76.9)	92(75.4)

적절한 연구방법과 통계적인 적용 및 해석은 의학적인 논문에서 점점 필수적인 요건이 되어 가고 있다.

계명의대논문집에서는 분석적인 연구방법에 속하는 사례-비교군 연구, 코호트 연구, 임상시험이 각각 9.3%, 1.9%, 6.4%가 있었으나, 엄격한 의미에서 연구방법론에 적합한 디자인(design), 실행과 통계적인 분석 및 해석(DerSimonian 등, 1982; Pocock, 1983; Meinert, 1986; Knapp과 Miller, 1992)을 한 논문은 거의 없었다고 할 수 있다. 전세계적인 추세가 단순한 임상적 고찰이나 횡단면적인 연구에서 벗어나 분석적인 연구, 특히 코호트 연구와 임상시험으로 이행된 단계이므로 이를 다룬 임상역학(clinical epidemiology)에 대한 개념확산이 되어 원인연구(etiology study)에 주력해야 함이 시급함을 알 수 있다. 임상적인 고찰을 하는 이유는 보기 드문 질병에 대해서 기술통계를 보아 가설을 설정하고(hypothesis generation) 원인연구를 시행할 전단계의 연구(precursor study)로서 의미를 지니고 있으나(Fletcher 등, 1988) 임상적 고찰에서 점검해 보아야 할 새로운 가설을 제시한 논문은 거의 없었다. 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 확대해석해서는 곤란할 것이다. 동물과 인간의 실험연구가 상당수 있었는데, 실험방법론에 입각한 실험과 통계적인 분석을 하게 되면 적은 비용으로 최대한의 가설검정을 할 수 있어 경제적인 연구가 될 수 있으므로(Fleiss, 1986; Kerlinger, 1986; Christensen, 1988) 이에 대한 주의가 요구된다. 이형기와 안윤옥(1991)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, 10년간 추이에서도 연구방법론의 변화양상이 뚜렷하지 않았다고 보고한 바 있다. 그러나 Feinstein(1978)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(1982)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 기술을 했고 추적손실은 79%, 치료순응도는 64%, 등록기준(eligibility criteria)은 37%, 통계적인 검정력은 12%에서 기술을 했다고 보고하였다.

이형기와 안윤옥(1991)은 분석한 총 382편 중 통계적인 처리방법을 밝히지 않은 논문이 117편으로서 통계적인 방법을 사용한 221편을 분모로 하면 52.9%나 되며, 통계처리기법을 밝히지 않은 경우를 제외하면 t 검정이 11.8%, Pearson의 상관분석이 11.0%, 분할표(contingency table, X^2 검정)가 9.4%로 대부분 간단한 단일분석이 대다수를 차지하고 있음을 보고하였다. 1989년의 대한예방의학회지에 게재된 55편의 논문에서 기술통계가 22편, t 검정 22편, X^2 검정 15편, 회귀분석 10편, 상관분석 8편, 분산분석 12편, 비모수검정 4편, 기타가 14편이었다(직업병연구소, 1990). 국외의 예로는 Felson 등(1984)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 보고 하였다. 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(1991)이 1990년 New England Journal of Medicine지에 발행된 순서대로 평가한 100편의 논문을 동일한 잡지의 1978-1979년에 게재된 논문을 분석한 Emerson과 Colditz (1983)의 결과와 비교를 했을 때, 기술통계만을 사용한 논문이 27%에서 11%로 감소했으나 간단한 단일변수기법은 거의 변화가 없었다고 보고하였다. 그러나 선형회귀분석과 비모수검정법(non-parametric methods)은 거의 2배의 증가를 보였으며, 좀 더 복잡한 다변수분석은 극적인 증가를 보였다. 큰 폭으로 증가한 기법은 생존분석법(survival analysis)으로서 100편 중 27편에서 이 방법을 사용했는데 대부분 다중로지스틱회귀분석(multiple logistic regression)과 Cox 회귀분석법(proportinal hazard method)의 증가였다고 보고하였다. 논문 편수당 서로 다른 통계적인 기법을 사용한 횟수가 증가하는 경향을 보였다고 하였다. 분산분석의 사용은 비교적 제한된 상황인데, 암관련 원저에서 2%정도, 정신과 잡지에서는 거의 10%였다(Hokanson 등, 1986). 계명의대논문집에서는 통계적인 방법을 사용하지 않았거나 또는 기술통계만을 사용한 논문이 총 36.1%였으나 매 기간마다 점차적으로 감소하여 82-83년에 68.2%였던 것이 90-91년에는 20.8%가 되었다. 그러나 결과 또는 표에 p 값만 제시해두고(orphan p) (Dawson-Saunders와 Trapp, 1990) 어떤 통계적인 방법을 사용했는지 알 수 없는 경우가 25편으로 10.4%나 되었으며, 더욱 문제인 것은 연도에 따라 감소하는 추세를 보이지 않는다는 점이다. 심지어는 "유의했다"고만 기술된 논문도 있었다. 최소한 재료 및 방법에 어떤 목적으로 어떤 통계기법을 사용하였다는 점은 기술해 두어야 할 것이다. 역학적인 방법을 사용한 논문은 총 6편으로 2.5%를 나타내었는데, 유병률과 발생률 그리고 민감도 연구 등이었으나, 유병률과 발생률을 혼동한 경우가 있었다. 다변수분석(multivariable analysis)으로는 다중선형회귀분석이 1편 있을 뿐이었다. 자료분석의 전세계적인 추세가 단일분석(univariate analysis), 층화분석(stratified analysis)의 단계를 지나 현재는 다변수분석의 단계에 접어들었으므로(Hanley, 1983; Dawson-Saunders와 Trapp, 1990) 이의 적용이 시급함을 알 수 있다. 특히, 여러 변수에 대해서 2그룹간에 단일변수로 유의성 검정을 하게 되면 변수들간의 상관성을 고려하지 못하게 되므로 통계적인 제1종오차의 가능성이 커지므로 다변수분석을 권고하고 있는 상황이다(Cupples 등, 1984). 그러나 다변수분석은 그 자체가 까다로운 기본적인 가정과 통계학적인 모델링(statistical modelling)의 문제를 가지고 있으므로 적용시 주의를 요한다(Hanley, 1983).

우리나라에서 통계적인 적용과 해석의 문제는 의학적인 분야에서만 문제가 되는 것이 아니라 전 학문영역에서 공통적으로 볼 수 있는 현상이다. 최종후 등(1990)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 보고한 바 있다. 단계별로 제시한 문제점들을 보면, 먼저 연구설계과정에서는 연구 대상에 대한 사전 탐사의 부족, 대표성의 고려 및 표본 크기에 대한 언급의 부족, 자료탐사의 무시, 실험설계의 미숙함, 측정의 문제 등이 있었으며, 통계적인 추론의 단계에서는 적절한

통계적 기법 선택의 문제와 적용절차상의 문제, 유의수준의 적용문제, 활용한 통계패키지의 구체적인 언급의 부족, 연구 과제의 통계적인 형식화의 고려부족 등이었다. 마지막 단계인 결론 도출 과정에서는 연구가설에 대한 결론의 기술문제가 가장 문제가 되었는데 결론의 서술시 이를 정당화시키는데 필요한 통계량이나 p 값이 제시되고, 이를 바탕으로 엄격한 서술이 이루어져야 하나, 대부분의 논문이 이 점을 소홀히 했으며, 특히 결론 기술을 너무 단정적으로 표현하는 것도 문제점으로 지적한 바 있다. 1989년 예방의학회지에 게재된 37편의 통계적인 방법론의 타당도 평가(직업병연구소, 1990)에서 역시 3단계로 나누어서 문제점을 제시한 바 있다. 이형기와 안윤옥(1991)이 1980년 1월부터 1989년 12월까지 대한의학협회지에 발행된 원저 382편을 분석했을 때, 통계적인 타당성 평가의 어느 항목이라도 적용이 가능한 총 297편의 원저 중 290편(97.6%)에서 최소한 하나 이상의 오류를 범했음을 보고하였다. 특히, 검정력이나 신뢰구간을 제시하지 않았으며 가설검정을 중복시행한 점 등이 많았음을 들었다. 앞의 저자들이 평가한 내용이 계명의대논문집에서도 그대로 노정되고 있는 문제점이라 할 수 있다.

이 연구에서 통계적인 기법의 적용이 잘못된 예가 상당한 비율임을 알 수 있었는데, 그 예 중 가장 흔한 것이 반복측정을 했을 때 처음 측정한 기초측정치(baseline measure)를 대조군으로 두고 각 시간별 측정치를 짝비교 t 검정(paired t-test)을 반복해서 실시한 경우와 3그룹 이상의 평균치 비교시에 독립 t 검정(independent t-test)을 반복해서 실시한 경우였다. 이러한 다중 t 검정(multiple t-test)시 문제는 통계적인 검정시에 항상 범하게 되는 제1종 오차(type I error)가 커져서 실제로는 유의하지 않은 그룹간에 유의한 것으로 나타날 가능성이 커진다는 것이며(error inflation), 이를 방지하기 위해서는 분산분석을 실시한 후 연구목적에 적합한 다중비교(multiple comparisons)를 하는 것이 권장할 방법이다(Smith 등, 1987; Shott, 1990; Bailar와 Mostellar, 1992). X^2 검정시 기대빈도가 5이하인 cell이 2×2 분할표에서 한 개 이상이 있거나 rXc 분할표에서는 약 25%를 넘게 되면 더이상 X^2 분포를 따른다고 볼 수 없으므로 이의 적용이 불가능한데(Armitage와 Berry, 1987) 상당수의 검정에서 이에 대한 기술이 없었으며, 실제 적용이 잘못된 예도 많았다. 이때는 2×2 분할표이면 Fisher의 정확확률검정(Fisher's exact test)을 해야 하며, rXc 분할표에서는 임상적인 의미에 손상을 주지 않는 범위에서 범주를 모아서(collapse) 기대빈도가 5이하인 cells의 비율을 25%이하로 해주는 것이 좋다(Hill과 Hill, 1991). 통계적인 방법을 기술시에 t 검정시에 독립(independent)과 짝비교(paired)를 구분하지 않은 논문이 많았다. 회귀분석(linear regression)과 상관분석(correlation analysis)을 혼동해서 사용한 경우가 많았으며, 상관분석도 자료의 성상에 따라 Pearson과 Spearman의 방법을 구분해서 사용해야 하나 이를 구분하지 않은 경우가 대부분이었다. 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(Petrie, 1987). 그러나 위와 같이 모수적인 방법을 적용시킬 수 없을 때 적용시켜서 얻을 수 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(Dawson-Saunders와 Trapp, 1990) 비모수검정법 사용에 대한 인식이 달라져야 할 것이다. 역학적인 방법을 사용한 논문은 총 6편으로 2.5%를 나타내었는데, 유병률과 발생률, 민감도 연구 등이었다. 유병률과 발생률을 혼동한 경우가 있었다. 민감도 연구와 같은 진단검사의 연구는 주의해야 할 사항과 연구계획시에 고려해야 하는 점이 많이 있으며 분석방법도 민감도, 특이도외에 양성 및 음성 예측도, ROC 곡선(receiver operator characteristic curve), likelihood ratio, 민감도 분석(sensitivity analysis) 등이 있으므로 상당히 철저한 계획과 분석이 필요한 연구임을 인식해야 할 것이다. 특히, 진단방법을 적용하는 대상의 질병의 유병률(prior probability)에 따라 결과가 달라질 수 있음을 알아야 한다(Fletcher 등, 1988; Hulley와 Cummings, 1988; Sackett 등, 1991). 역학적인 통계기법의 하나인 비교위험도(relative risk) 또는 대응

비(odds ratio)를 사용한 논문은 거의 없었으며, 혼란변수(confounding variables)에 대한 개념과 통제의 방법에 대한 기술 역시 거의 없어서 임상역학에 대한 개념정립이 시급함을 보여 주었다.

재료 및 방법에서는 최소한 사용한 통계적인 기법, 사용에 따른 기본적인 가정이 충족되는지의 여부, 통계적인 유의성을 결정하는 유의수준 등에 대한 기술이 되어야 한다. 계명의대논문집에서는 사용된 통계적인 기법을 기술하지 않은 비율이 연도별로 줄고 있으나 90-91년에도 여전히 23.1%라는 높은 비율을 보였으며, 사용된 통계 기법의 기본적인 가정이 논문의 목적과 자료의 성상에 적합한지를 점검한 논문은 총 3편으로 2.5%에 지나지 않아 심각한 문제가 되었다. 통계기법의 기본적인 가정에 어긋남에도 불구하고 적합하지 않은 통계기법을 사용하게 되면 잘못된 결과와 이에 따른 해석이 도출되어 논문의 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 계산해보고(Kraemer, 1988; Lemeshow 등, 1990) 이에 맞추어 대상자를 얻은 논문은 거의 없었다. 또, 연구결과가 기대했던 결과를 나타내지 못했을 때(negative findings) 시행된 연구의 표본크기에 따른 통계적인 검정력(power)에 대한 점검을 해보아 통계적인 제2종오차(type Ⅱ error)의 정도를 고려해 보아야 하나(Freiman 등, 1978; Pocock, 1983; Hauck과 Anderson, 1986) 이를 고찰한 논문 역시 거의 없었다. 현재는 개인용 컴퓨터를 이용해서 대부분의 연구방법과 통계기법에 대해 표본크기와 검정력을 계산해주는 프로그램이 있으므로 (BMDP, 1991) 연구를 시작하기 전에 반드시 표본크기를 계산한 후 대상자 선정과 등록의 단계에 들어가야 할 것이다. 이는 특히 임상시험(clinical trial)일 경우 중요한 의미를 지닌다(Meinert, 1986). 그리고 통계적인 유의성(statistically significance)이 항상 임상적인 유의성(clinical significance)을 의미하는 것이 아니므로 통계적인 유의성이 연구결과를 해석하는데 있어 절대적인 기준이 되지 못함을 알아야 한다. 특히, 점추정(point estimation)으로서의 p 값은 큰 의미를 지니지 못하나(Diamond와 Forrester, 1983; Woolson과 Kleinman, 1989; Goodman, 1993), 최근들어 통계적인 기법이 우리나라 의학계에 도입이 되면서 연구자들이 너무 p 값에 의존하는 경향을 보이고 있다. 신뢰구간은 독자들에게 연구에서 제시된 추정치가 변이를 가지고 있어 만약에 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 (Simon, 1986; Dawson-Saunders와 Trapp, 1990). 국외의 저명한 의학잡지에서는 통계적인 검정시 p 값에 너무 의존하는 경향에서 벗어나 신뢰구간(confidence intervals)의 추정으로 방향을 전환하고 있는 추세이다(Campbell과 Machin, 1993). British Medical Journal의 편집자들은 신뢰구간이 적절할 경우에는 가설검정 대신에 신뢰구간을 제시하도록 하는 방침을 정해두고 있다(Gardner와 Altman, 1986).

이 연구에서 가장 문제를 보인 점검항목은 표본표집에 관한 것이다. 대표성을 지닌 표집법을 사용한 논문이 1편에 지나지 않았다. 그리고 표본 표집에 대한 기술이 없거나 잘 모를 경우가 28편으로 19.6%나 되었으나 연도별로는 감소추세를 보였다. 문제가 되는 논문은 1개 대학병원에 내원한 환자를 대상으로 연간 발생률(annual incidence) 등의 통계를 본 것이다. 또, 표본크기를 제시하지 않은 논문도 있었다. 대부분의 의학문헌에서 대상이 되는 환자들은 편의가 게재된 표본(biased samples)이며, 특히 종합병원이나 대학병원에서 보는 환자들은 전체 대상 환자의 극히 일부분에 지나지 않는다. 이러한 표집법은 환자를 대상으로 하는 의학연구에서는 그 자체로는 문제가 되지 않으나 이들을 대상으로 얻은 결과를 확대해석(generalization)하는 것이 문제가 된다(Fletcher 등, 1988). 몇 편의 논문에서는 병원에 내원한 사람들을 대상으로 하고서 제목에 "한국인의" 식으로 기술되어 있었다. 이는 연구집단(study population), 표본집단(sampled population), 대상집단(target population)에 대한 개념이 없는데 기인한다고 볼 수 있다(Knapp과 Miller, 1992). 대상자 선택(subject selection)과 연관된 편의로는 Neyman bias, Berkson's fallacy, nonresponse bias, membership bias, procedure selection bias 등이 있

으므로 결과해석에 특히 주의를 요한다(Dawson-Saunders와 Trapp, 1990).

이러한 통계적인 적용상의 문제점들을 해결해서 타당도를 향상시키기 위해서 대학과 대학원 과정에서 올바른 통계학적인 교육이 선행되어야 하고, 게재를 위해 제출된 논문에 대해서 전문가들이 사독과정을 거쳐서 조언을 해줄 수 있는 체계(referee system)의 확립 그리고 통계적인 타당도는 연구의 디자인과 불가분의 관계를 가지고 있으므로 연구를 디자인하는 단계에서부터 통계 전문가들과 협력해서 연구종료시에 타당한 통계적인 적용과 해석을 할 수 있도록 도와줄 수 있는 통계상담(statistical counselling) 서어비스체제의 완비와 이를 활용하고자 하는 연구자들의 의지 그리고 논문집에 독자의 반론을 실을 수 있는 지면의 마련 등이 제시되고 있다(Glantz, 1980; 최종후 등, 1990; Appleton, 1990; 이형기와 안윤옥, 1991). 투고규정에 최소한 형식적으로 갖추어야 할 통계학적인 기준을 제시하고 이를 어겼을 경우에는 다시 수정하도록 하는 제도 역시 고려해 볼만하다.

이 조사의 제한점으로는 대학잡지에 실리는 논문은 선택적일 경우가 많아서 국내의 다른 학회논문집과의 비교보다는 다른 대학논문집과의 비교가 더 타당하나 이에 대한 자료가 부족해서 비교가 힘들었다는 점이다. 그리고 최종후 등(1990)의 점검사항에 따라 연구설계과정, 통계적인 추론, 결론도출과정에 대해 자세하게 세분해서 평가를 하지 않고 현실적인 문제점으로 해서 객관적이고 형식적인 부분에 대해서만 평가를 했다는 점이다.

요 약

계명의대논문집의 통계적인 타당도를 제고할 기초자료를 제공하기 위해 1982-1991년간 발행된 논문 204편을 대상으로 통계학적인 타당도를 분석 고찰하였다. 개발된 타당도의 점검표는 계명대학교 의과대학 예방의학교실과 외부 통계학과의 대학원생들에 의해 평가되었다. 횡단면적인 연구가 81편(39.7%)으로 가장 많았으며 동물실험은 61편으로 30.0%를 차지했다. 코호트 연구는 4편(1.9%)에 지나지 않았으며 88-89년부터 나타나기 시작하였다. 통계적인 방법을 사용하지 않았거나 또는 기술통계만을 사용한 논문이 87건(36.1%)이었으며 연도별 감소하였다. t 검정은 68건(28.2%)이였다. 비모수검정과 생존분석 그리고 다중회귀분석은 각각 1건에 지나지 않았다. p값만 제시하고 사용한 통계기법을 알 수 없었던 논문이 25건(10.4%)이나 되었으며 기간에 따라 별 변동이 없었다. 대표성을 지닌 표본 표집법을 사용한 논문은 1편에 지나지 않았고, 표집법에 대한 기술이 없거나 잘 모를 경우가 28편으로 19.6%를 차지했으나 연도에 따라 감소하는 추세를 보였다. 재료 및 방법에 사용한 통계적인 기법을 기술한 논문은 73편(59.8%)이었으며 연도에 따라 증가하는 추세를 보였다. 기본적인 가정이 충족되었는지를 점검한 논문은 3편으로 2.5%에 지나지 않았다. 통계적인 유의수준을 제시한 논문은 12편(9.8%)이었으나 연도에 따라 오히려 감소하는 추세를 보였다. 통계적인 기법이 적절하게 된 논문은 30편(24.6%)이었으며 연도별로 증가하는 경향을 보였다.

이러한 결과는 계명의대논문집에서 통계학적인 적용과 해석에 문제의 소지가 많으며, 통계적인 타당동의 향상을 위해서는 논문 게재를 원하는 저자와 편집자 양측 모두에서 상당한 노력에 필요함을 시사해주는 것이다.

참 고 문 헌

근로복지공사 중앙병원부설 직업병연구소: 산업보건과 통계적 방법론-예방의학회지에 발표된 연구논문의 통계적 방법론적 타당성 평가. 직연보 23-90-12, 서울, 1990.

안윤옥, 고응린: 자료처리과정에 대한 통계학적 검토-일부 의학잡지에 게재된 논문예를 중심으로. 예방의학회지 1973; 6: 81-85.

안윤옥, 이형기: 의학에서의 연구방법론. 한국역학회지 1990; 12: 107-114.

안윤옥, 이형기: 의학연구논문의 방법론 및 통계처리기법의 타당성 평가를 위한 점검표 개발. 한국의학교육 1991; 319-35.

윤기중, 안윤기, 김병수: 통계의 오용과 효율적 이용에 관한 연구. 산업과 경영 1987; 24: 3-37.

이형기, 안윤옥: 1980년대에 발표된 국내 의학연구논문의 방법론 및 통계처리기법의 타당성에 관한 평가연구. 한국의학교육 1991; 3: 52-69.

최종후, 김기목, 김기영등: 과학학술지에 나타난 통계적 기법활용의 타당성 평가. 응용통계 1990; 5: 1-16.

최종후, 이재창: 학술논문과 통계적 기법, 고려대학교 통계연구소, 통계분석강의총서 12. 서울, 자유아카데미, 1990.

Altman DG: Statistics in medical journals: Developments in the 1980s. *Stat Med* 1991; 10: 1897-1913.

Andersen B: *Methodological Errors in Medical Research: An incomplete catalogue*. London, Blackwell Scientific Publications, 1990, pp 1-24.

Appleton DR: What statistics should we teach medical undergraduates and graduates? *Stat Med* 1990; 9: 1013-1021.

Armitage P, Berry G: *Statistical Methods in Medical Research*, ed 2. London, Blackwell Scientific Publications, 1987, pp 371-386.

Bailar Ⅲ JC, Mosteller F: *Medical uses of Statistics*, ed 2. Boston, NEJM Books, 1992, pp 233-257.

BMDP: *SOLO Statistical System: Power analysis*. Los Angeles, BMDP Statistical Software, Inc., 1991.

Campbell MJ, Machin D: *Medical Statistics: A commonsense Approach*. John Wiley & Sons, Chichester, 1993, p 85.

Chalmers I: Misconduct in medical research. *Br Med J* 1988; 298(6668): 256.

Christensen LB: *Experimental Methodology*, ed 4. Boston, Allyn and Bacon, Inc., 1988.

Cupples LA, Heeren T, Schatzkin A, et al: Multiple testing of hypotheses in comparing two groups. *Ann Int Med* 1984; 100: 122-129.

Dawson-Saunders B, Trapp RG: *Basic and Clinical Biostatistics*. East Norwalk, Prentice-Hall International Inc, 1990, pp 64-98, 110-111, 207-228, 264-275.

DerSimonian R, Charette LJ, McPeek B, et al: Reporting on methods in clinical tirals. *N Engl J Med* 1982; 306: 1332-1337.

Diamond GA, Forrester JS: Clinical trials and statistical verdicts: Probable grounds for appeal. *Ann Int Med* 1983; 98: 385-394.

Dunn HL: Application of statistical methods in physiology. *Physiol Rev* 1929; 9: 275-398.

Emerson JD, Colditz GA: Use of statistical analysis in the New England Journal of Medicine. *N Engl J Med* 1983; 309: 709-713.

Evans M, Pollock AV: A score system for evaluating random control trials of prophylaxis of abdominal surgical wound infection. *Br J Surg* 1985; 72: 256-260.

Evans M, Pollock AV: Inadequacy of published random control trials of antibacterial propylaxis in colorectal surgery. *Dis Colon Rectum* 1987; 30: 743-746.

Feinstein AR: Clinical biostatistics: A survey of the research architecture used for publications in general medical journals. *Clin Pharmacol Ther* 1978; 23: 117-125.

Felson DT, Cupples LA, Meenan RF: Misuse of statistical methods in arthritis and rheumatism. *Arthritis Rheum* 1984; 27: 1018-1022.

Fleiss JL: *The Design and Analysis of Clinical Experiments*. New York, John Wiley & Sons, 1986.

Fletcher RH, Fletcher SW, Wagner EH: *Clinical Epidemiology: The Essentials*, ed 2. Baltimore, Williams & Wilkins, 1988, pp 188-207.

Freiman JA, Thomas AB, Chalmers C, et al: The importance of beta, 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: 690-694.

Gardner MJ: Understanding and presenting variation. *Lancet* 1975:230-231.

Gardner MJ, Altman DG: Confidence intervals rather than P values: Estimation rather than hypothesis testing. *Br Med J* 1986; 292: 746-750.

Glantz SA: Biostatistics: How to detect, correct and prevent errors in medical literature. *Circulation* 1980; 61, 62.

Goodman SN: P values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *Am J Epidemiol* 1993:137:485-496.

Greenwood M: What is wrong with the medical curriculum? *Lancet* 1932; 1: 1269-1270.

Hanley JA: Appropriate uses of multivariate analysis. *Ann Rev Public Health* 1983; 4: 155-180.

Hauck WW, Anderson S: A proposal for interpreting and reporting negative studies. *Stat Med* 1986; 5: 203-209.

Hebert JR, Miller DR: Methodologic considerations of investigating the diet-cancer link. *Am J Clin Nutr* 1988; 47: 1068-1077.

Hill AB, Hill ID: *Bradford Hill's Principles of Medical Statistics*, ed 12. London, Edward Arnold, 1991, pp 136-147.

Hokanson JA, Luttman DJ, Weiss GB: Frequency and diversity of use of statistical techniques in oncology journals. *Cancer Treat Rep* 1986; 70: 589-594.

Hulley SB, Cummings SR: *Designing Clinical Research: An epidemiologic approach.* Baltimore, Williams & Wilkins, 1988, pp 85-97.

Kerlinger FN: *Foundations of Behavioral Research*, ed 3. New York, Holt, Rinehart and Winston, 1986, pp 279-346.

Knapp RG, Miller MC: *Clinical Epidemiology and Biostatistics.* Malvern, Harwal Publishing Company, 1992, pp 93-130.

Kraemer HC: Sample size: When is enough? *Am J Med Sci* 1988; 296: 360-363.

Lemeshow S, Hosmer DW, Klar J, et al: Adequacy of Sample Size in Health Studies. John Wiley & Sons, Chichester, 1990, pp 1-48.

Meinert CL: *Clinical trials: Design, Conduct, and Analysis.* New York, Oxford University Press, 1986, pp 63-89.

Petrie A: *Lecture Notes on Medical Statistics*, ed 2. Oxford, Blackwell Scientific Publications, 1987, pp 171-186.

Pocock SJ: *Clinical Trials: A Practical Approach.* Chichester, John Wiley & Sons, 1983, pp 123-141.

Pocock SJ: Statistics in medicine:current issues in the design and interpretation of clinical trials. *Br Med J* 1985; 290: 39-42.

Sackett DL, Haynes RB, Guyatt GH, et al: *Clinical Epidemiology: A Basic Science for Clinical Medicine*, ed 2. Boston, Little, Brown and Company, 1991, pp 69-152.

Schoolman HM, Berktel JM, Best WB, et al: Statistics in medical research: Principles versus practices. *J Lab Clini Med* 1968; 71:357-367.

Schor S, Karten I: Statistical evaluation of medical journal manuscripts. *J Am Med Assoc* 1966; 195: 1123-1128.

Shott S: *Statistics for Health Professionals.* Philadelphia, WB Saunders Company, 1990, pp 145-166.

Simon R: Confidence intervals for reporting results of clinical trials. *Ann Int Med* 1986; 105: 429-435.

Smith DG, Clemens J, Crede W, et al: Impact of multiple comparisons in randomized clinical trials. *Am J Med* 1987; 83: 545-550.

Thoron MD, Pulliam CC, Symons MJ, et al: Statistical and research quality of the medical and pharmacy literature. *Am J Hospital Pharma* 1985; 42: 1077-1082.

Williamson JW, Goldschmidt PG, Colton T: The quality of medical literature: An analysis of validation assessment, Bailar JC, Mostellar F, (eds): *in Medical Uses of Statistics.* Massachusetts Medical Society, 1986.

Woolson RF, Kleinman JC: Perspectives on statistical significance testing. *Ann Rev Public Health* 1989; 10: 423-440.

＝Abstract＝

Review of the validity of statistical methods in the Keimyung University Medical Journal, 1982－1991

Nung Ki Yoon, MD; Choong Won Lee, MD; Suk Kwon Suh, MD; Jong Won Park, MD

Department of Preventive Medicine,
Keimyung University School of Medicine, Taegu, Korea

Authors reviewed the 204 original articles published in the Keimyung University Medical Journal during 1982－1991 to give a status report of validity of statistical methods for the purpose of improving the validity of statistical applications. Checklists were assessed by the independent reviewers from the department of statistics of a graduate school. Number of cross-sectional study was 81(39.7%) and that of animal study was 61(30.0%). Cohort study was merely 4(1.9%) which appeared first during 1988-1989. Eighty-seven articles used no statistical method or descriptive statistics only, frequency of which decreased yearly and 68 articles(28.2%) used t-test. Nonparametric test, survival analysis and multiple linear regression were used once respectively. Orphan p where no statistical method had been specified and only the p value given was presented in 25 article(10.4%) which showed little fluctuations yearly. Representative sampling method was employed only in one article and no description of sampling method was noted in 28 articles(19.6%), but the yearly frequency was decreased steadily. Descriptions of statistical methods used in materials and methods was noted in 73 articles(59.8%), number of which appeared to increase yearly. Basic assumptions of the statistical methods used were discussed in 3 articles only(2.5%). Twelve articles(9.8%) described the significance level declared statistically significant, showing yearly decreasing frequency. Applications of the statistical methods appeared to be appropriate in 30 articles(24.6%), frequency of which showing increasing yearly trends. These results suggest that the validity of statistical methods used appears to be seriously compromised in this period and has much to be done to improve the current situations.

Key Words: Statistical validity