예방의학회지 : 제30권 제4호
Korean J. of Preventive Medicine
Vol. 30, No. 4, December. 1997

로그선형모형을 이용한 복수 평가자들간의 신뢰도에 관한 연구

박병주[1], 이성임[2], 이영조[2], 김동현[3], 권호장[4], 배종면[5], 신명희[6], 하미나[4], 한상환[7]

서울대학교 의과대학 예방의학교실[1], 자연과학대학 통계학과[2], 한림의대 사회의학교실[3],
단국의대 예방의학교실[4], 충북의대 예방의학교실[5], 성균관의대 예방의학교실[6], 길의료원[7]

= Abstract =

Reliability for Multiple Reviewers by using Loglinear Models

Byung-Joo Park[1], Sung-Im Lee[2], Youngjo Lee[2], Dong-Hyun Kim[3], Ho-Jang Kwon[4],
Jong-Myon Bae[5], Myung-Hee Shin[6], Mi-Na Ha[4], Sang-Whan Han[7]

Department of Preventive Medicine[1], Department of Statistics[2], Seoul National University, Department of Social Medicine[3], Hallym University, Department of Preventive Medicine[4], Dankuk University, Department of Preventive Medicine[5], ChungBuk University, Department of Preventive Medicine[6], SungKyunKwan University, Institute of Occupational Health, Gil Medical Center[7]

To guarantee the inter-reviewer reliability is very important in evaluating the quality of large number of clinical research papers by multiple reviewers. We cannot find reports on statistical methods for evaluating reliability for multiple raters in clinical research field. The purpose of this paper is to introduce the statistical methods focused on kappa statistic and five kinds of loglinear models for, which can be applied when evaluating the reliability of multiple raters. We have applied these methods to the result of a project, in which seven reviewers have evaluated the quality of 33 papers with regard to four aspects of paper contents including study hypothesis, study design, study population, study method, data analysis and interpretation. Among the five loglinear models including Symmetry model, Conditional symmetry model, Quasi-symmetry model, Independence model, and Quasi-independence model, Quasi-symmetry model shows the best model of fitting. And the level of reliability among seven reviewers revealed to be acceptable as meaningful.

Key words : reliability, multiple reviewers, kappa, loglinear model

* 본 연구는 한국과학재단 특정기초연구비(960701-01-01-3)를 지원받아 수행되었음.

Ⅰ. 서 론

국내에서 임상의학연구는 의과대학과 임상의학연구자의 수적 증가에도 불구하고, 기초의학연구에 비해 상대적으로 그 질적가치를 인정받지 못하고 있는 실정이다. 그 이유중에 한가지로, 최근에 이르기까지 임상의학연구가 과학적이고 체계적으로 수행되고 있지 못하다는 점이 지적된 바 있다(박병주, 1994).

이에 국내 임상의학연구의 질적 수준을 객관적으로 평가해 보려는 시도가 요구되었고, 본 연구진들은 국내 2개 의과대학에서 1987년과 1992년의 2개년에 걸쳐 발표된 971편의 임상의학 연구논문들과 같은 시기에 미국의사협회의 정기학술지인 Journal of the American Medical Association(JAMA)에 실린 225편의 임상의학연구논문들을 대상으로 그 질적 수준을 비교 평가하려는 연구를 계획하였다. 이를 위하여 7인의 평가자들이 연구대상으로 선정된 1,196편의 연구논문을 개인별로 할당하여 독립적으로 평가하였다. 이와 같은 다수의 평가자들이 공동으로 수행하는 논문평가작업에 있어서 평가자들간의 신뢰도가 일정수준 이상으로 유지되어야 한다는 것은 매우 중요하다. 환언하자면, 동일한 논문에 대한 평가자들간의 평가결과에 상당한 차이가 있다면 이러한 공동작업은 그 의미를 상실하게 된다고 할 수 있다. 실질적으로 평가자들간의 신뢰도를 높히기 위하여 연구책임자가 나머지 평가자들에게 관련문헌들을 이용하여 일정기간동안 논문평가에 관한 집중교육을 실시하였고, 평가대상논문들의 연구가설, 설계형태, 연구대상선정, 연구방법, 자료의 통계적 분석 및 결과해석 등에 따른 분류별로 표준화된 논문평가표(Article check list; 부록 참조)를 미리 작성하여 사용함으로써 복수 평가자들간의 평가기준을 표준화하기 위하여 노력하였다.

이러한 노력으로 평가자들간의 평가기준이 과연 일정 수준이상의 신뢰도를 갖추게 되었는지를 파악하기 위하여, 실제 평가작업 대상인 1,196편의 논문중 무작위로 33편의 논문을 선정하여 이들을 대상으로 공통된 항목들에 관한 평가결과를 카파(kappa)척도와 로그선형모형을 이용하여 통계적으로 비교분석하였다.

Ⅱ. 연구 방법

1. 복수평가자들의 일치도 평가를 위한 모델의 적용

복수평가자들은 서울대학교 의과대학 예방의학교실에 근무중인 연구책임자와 6명의 전공의들로 구성되었다. 신뢰도를 평가하기 위한 연구대상 논문은 전체 평가대상 논문 1,196편의 임상연구논문중에서 무작위로 33편을 선정하였다. 선정된 논문들에 대하여 연구가설의 명확성(논문평가표에서 1-3까지의 항목), 연구대상의 적절성(논문평가표에서 4-5까지의 항목), 자료수집과정의 신뢰성(논문평가표에서 6-7까지의 항목), 및 자료의 통계적 분석 및 해석의 적절성(논문평가표에서 8-11까지의 항목) 등 4가지 공통된 영역을 평가하도록 하였다. 논문평가표에서 각 항목의 내용은 부록에 첨부하였다. 각 항목에는 순위척도로 점수가 주어져 있다.

4가지 영역의 점수는 각 항목들의 점수 합계로써 나타내고, 다시 이들 점수를 가장 신뢰성 있는 평가자가 각 영역에 따라 적절한 순위척도를 이용하여 재분류하였다. 첫째, 연구가설의 명확성은 (1)명확하지 않다 (2)비교적 명확하다 (3)매우 명확하다; 둘째, 연구대상의 적절성은 (1)적절하지 않다 (2)적절하다; 셋째, 자료수집과정에 대한 신뢰성은 (1)신뢰할 수 없다 (2)신뢰할 수 있다; 마지막으로 결과의 통계적 분석 및 해석의 적절성은 (1)매우 부족하다 (2)다소 부족하다 (3)보통이다 (4)충분하다로 나누었다.

주어진 범주로 재구성된 자료를 바탕으로 아래에서 소개한 통계적 방법을 이용하여 평가자들중 연구책임자와 나머지 평가자들간의 일치도를 측정하였다. 본 연구에서 연구책임자는 평가자들중 가장 신뢰성있는 연구자로서 연구책임자의 평가기준이 가장 객관적일 것으로 가정하였다. 각 영역별로 연구책임자와 나머지 6명의 평가자들과의 2×2 분할표 6개를 만들고, 연구

책임자(reviewer A)와 나머지 6명의 평가자들(reviewer B)간의 일치도 평가를 위하여 6개의 2×2 분할표의 도수를 모두 합하여 새로운 하나의 2×2 분할표(표1-4)를 재구성하였다.

2. 신뢰도를 평가한 통계적 방법

논문작성표를 통해 얻어진 평가 결과는 순위형 범주(ordinal category)를 값으로 하고 있으며 이는 분할표로 정리되어진다.

1) 일치도 (Measure of Agreement)

평가자들간의 평가에 대한 일치정도는 공동평가작업에 대한 신뢰도 수준을 반영하는 척도가 된다. 두 관찰자가 동일 대상을 순위척도(ordinal scale)로 분류하는 경우를 생각해보자. 예를 들어, 내과의사 2명이 환자의 건강상태를 네 가지 범주(매우 건강함, 건강함, 약간 불건강함, 매우 불건강함)로 분류한다면, 각 환자는 다음과 같이 4×4 분할표의 어떤 한 칸(cell)에 속하게 될 것이다. 이때 주대각(▨)은 두 관찰자의 분류가 일치함을 나타낸다. 모든 자료가 주대각에 몰려 있다면 두 평가자의 평가는 어느 정도 객관성을 띤다고 할 수 있다. 즉, 평가자들의 평가는 객관적이라 할 수 있어 신뢰할 수 있다. 분류행위란 주관적인 영향을 많이 받으므로, 그 분류의 신뢰도를 검토해보는 것은 매우 중요하다. 이때 신뢰도는 두 관찰자들의 분류결과가 일치하는가를 평가함으로써 가늠해 볼 수 있는데, 본 연구에서는 카파(kappa)라는 척도를 사용하였다.

2) 카파 (Kappa)

이제 π_{ij}를 첫 번째 관찰자가 i번째 범주, 두 번째 관찰자가 j번째 범주로 분류할 확률(probability of classification)이라고 정의하자. 이 경우

$$\prod_o = \sum \pi_{ii}$$

는 두 관찰자들이 서로 일치하는 확률로써 $\prod_o = 1$은 완전한 일치를 의미한다. 두 관찰자의 분류가 통계적으로 독립일 경우, 첫 번째 관찰자가 i번째 범주, 두 번째 관찰자도 i번째 범주를 택할 분류확률은 $\pi_{ii} = \pi_{i+}\ \pi_{+i}$가 된다. 이때 π_{i+}는 첫번째 관찰자가 i번째 범주를 택할 확률이고 π_{+i}는 두번째 관찰자가 j번째 범주를 택할 확률이다. 이 때의 일치 확률 (probability of agreement)은 다음과 같다.

$$\prod_e = \sum \pi_{i+}\ \pi_{+i}$$

따라서 분류가 통계적으로 독립적이라면 $\prod_o - \prod_e$는 관찰자들간의 일치도가 우연히 기대되는 것 이상의 초과분을 나타낸다. 관찰자들간의 일치도를 나타내는 대표적인 척도로 Cohen(1960)의 카파를 들 수 있다:

$$K = \frac{\sum \pi_{ii} - \sum \pi_{i+}\ \pi_{+i}}{1 - \sum \pi_{i+}\ \pi_{+i}} = \frac{\prod_o - \prod_e}{1 - \prod_e}$$

두 관찰자가 우연히 일치했을 경우 $\prod_o$와 $\prod_e$의 값이 동일하므로 카파값은 0에 해당되고, 두 관찰치가 완전히 일치했을 경우 $\prod_o = 1$이므로 카파값은 1이다. 따라서 일치도가 높으면 높을수록 카파값은 커지게 된다. 우연히 기대되는 정도보다 일치도가 더 약한 경

내과의사 A \ 내과의사 B	매우 건강함	건강함	약간 불건강함	매우 불건강함
매우 건강함	▨			
건강함		▨		
약간 불건강함			▨	
매우 불건강함				▨

Table 1. Some Loglinear Models of Evaluating Symmetry of I × I table

(1) Symmetry model	$\log m_{ij}=\mu+\lambda_i+\lambda_j+\lambda_{ij}$ 단, $\lambda_{ij}=\lambda_{ji}$
(2) Quasi-symmetry model	$\log m_{ij}=\mu+\lambda_i^X+\lambda_j^Y+\lambda_{ij}^{XY}$ 단, $i\neq j$일때 $\lambda_{ij}^{XY}=\lambda_{ji}^{XY}$
(3) Quasi-independence model	$\log m_{ij}=\mu+\lambda_i^X+\lambda_j^Y+\delta_i I(i=j)$
(4) Independence model	$\log m_{ij}=\mu+\lambda_i^X+\lambda_j^Y$
(5) Conditional symmetry model	$\log m_{ij}=\mu+\lambda_i+\lambda_j+\tau I(i<j)$ 단 $\lambda_{ij}=\lambda_{ji}$

우 (−)값이 나올 수 있으나 이런 경우는 매우 드물다.

주어진 범주가 3개 이상일 경우에 대해 카파의 표본값 $\hat{K}$는 점근적으로 정규분포를 따르며 그 분산은 다음과 같이 추정할 수 있다(Fleiss 등, 1969):

$$\hat{\sigma}^2(\hat{K})=\frac{1}{n}\left\{\frac{\hat{\prod}_o(1-\hat{\prod}_o)}{(1-\hat{\prod}_e)^2}+\frac{2(1-\hat{\prod}_o)\,[2\hat{\prod}_o\hat{\prod}_e-\sum\hat{\pi}_{ii}(\hat{\pi}_{i+}+\hat{\pi}_{+i})]}{(1-\hat{\prod}_e)^3}+\frac{(1-\hat{\prod}_o)^2\,[\sum\sum\hat{\pi}_{ij}(\hat{\pi}_{j+}+\hat{\pi}_{+j})^2-4\hat{\prod}_e^2]}{(1-\hat{\prod}_e)^4}\right\}$$

여기서 $\hat{\prod}_o=\sum\hat{\pi}_{ii}$이고 $\hat{\prod}_e=\sum\hat{\pi}_{i+}\hat{\pi}_{+i}$이다. 일반적으로 두 관찰자의 일치도가 우연히 기대되는 것에 지나지 않는 경우는 거의 없다. 따라서 귀무가설 H_0:K=0을 검정하는 것보다 카파값에 대한 신뢰구간을 추정하여 카파값의 크기를 평가하는 것이 더욱 의미있는 일이다.

3) I×I 분할표에 적용할 수 있는 로그선형모형

한편 두 평가자의 평가가 주대각이외에 넓게 분산되어 있다면, 두 사람의 평가결과는 신뢰할 수 없게 된다. 그러나 그러한 경우이더라도, 예를 들어, 관찰자 A가 관찰자 B보다 항상 한 단계 높게 분류할 경우, 일치도는 낮지만 연관도(measure of association)는 높게 나타날 수 있다. 연관도란 두 관찰자가 서로 독립적으로 평가하였는지를 알려주는 통계량이다. 연관도의 척도에는 감마(gamma), 켄달의 타우-b(Kendall's Tau-b), 스튜워트의 타우-c(Stuart's Tau-c), 및 소머의 D(Sommer's D) 등이 있다(김병수 등, 1987; 허명회, 1995; Agresti, 1984). 일치도가 크지 않을 경우는 우선적으로 평가자들의 평가가 상호 독립적인지 아닌지가 관심의 대상이 될 수 있다. 그러나 단순히 연관도만 가지고는 충분하지 않고, 실제 평가자들간의 상호평가경향을 설명하는 것이 필요하다. 즉 켄달의 타우-b값으로는 관찰자 A가 관찰자 B보다 높게 평가한 경우인지, 관찰자 B가 관찰자 A보다 높게 평가한 경우인지 알 수 없다. 따라서 이러한 상호적인 평가경향을 설명하기 위하여는 다양한 로그선형모형을 적합시켜 해석해 보아야 한다.

표 1에 소개한 모형들은 I×I 분할표를 분석하는데 사용되고 있는 대표적인 로그선형모형들이며, 이들 중 모형(5)는 특히 순위척도인 경우에 유용한 모형이다. 표 1에서 $m_{ij}=n\pi_{ij}$란 (i,j)칸의 기대도수를 의미하고 X는 $\{\pi_{ij}\}$분포에 의한 행의 관측값을, Y는 열의 관측값을 나타낸다.

로그선형모형을 이용하여 분석할 때 중요한 점은 대칭성에 대한 검토가 이루어져야 한다는 것이다. 즉, (i,j)칸의 확률이 분할표의 주대각에 대해 대칭으로 나타나는가와 주변 대칭성을 갖는지 알아보는 것이다. 부연하면, $i\neq j$인 경우 $\pi_{ij}=\pi_{ji}$를 만족하면 대칭성이 존재한다고 정의하고, i=1, …, I에 대해 $\pi_{i+}=\pi_{+i}$를 만족하면 주변 대칭성(marginal symmetry; 이를 주변 동질성(marginal homogeneity)으로 부르기도 한다)이 존재한다고 정의한다. 대칭성이 있으면 모든 i에 대해 $\pi_{i+}=\sum_j\pi_{ij}=\sum_j\pi_{ji}=\pi_{+i}$를 만족하므로 또한 주변 대칭성이 존재하게 된다. 그러나, I>2인 분할표의 경우에는 주변 대칭성이 존재하더라도 항상 대칭성이 만족되지는

않을 수도 있다.

모든 π_{ij}의 값이 0보다 클 때, 대칭성은 (1)의 모형으로 표현될 수 있다. 두 관찰자의 분류를 서로 동일한 인자 $\{\lambda_i\}$로 쓰고 있는 이 모형의 의미는 주대각에 대해 관찰값이 완벽히 대칭임을 시사한다. 다시 말해, (i,j)칸의 기대값과 (j,i)칸의 기대값이 같음을 의미한다. 그러나, 이 모형의 완벽한 대칭성을 만족하는 매우 제한적인 현상은 현실적으로 거의 존재하지 않기 때문에 대부분의 자료에 잘 적합하지 않는 것으로 알려져 있다. 모형(1)의 단점을 극복하기 위하여 Caussinus (1965)에 의해 제안된 모형(2)는 두 관찰자의 서로 다른 판단을 인정한다. 따라서 두 관찰자를 구별짓도록 λ_i^X와 λ_i^Y를 사용한다. 모형(1)은 모든 i에 대해 $\lambda_i^X=\lambda_i^Y$가 성립하는 특별한 경우이며, 이밖에 많은 유용한 모형들이 quasi-symmetry model의 특별한 경우로 제안되어 있다.

주대각이 독립적 모형이 예측하는 것 이상으로 큰 도수를 갖는 경우는 두 관찰자간의 일치도가 독립모형에서 예측한 것보다 더 높다는 것을 의미한다. 이같은 현상을 설명해주면서 동시에 주대각 이외의 경우에는 독립을 가정하고 있는 모형을 quasi-independence model(모형 (3))이라고 한다. 모형(4)는 두 관찰자의 반응이 서로 독립적임을 가정한 모형이며 이는 모형(2)가 $i \neq j$인 경우 λ_{ij}가 동일한 특별한 경우의 모형이 된다.

순위형 분류변수에서 모형(1)이 적절치 않을 경우, 모든 $i\langle j$에 대해 $\pi_{ij}\rangle\pi_{ji}$이거나 혹은 반대로 모든 $i\langle j$에 대해 $\pi_{ij}\langle\pi_{ji}$일 경우가 있다. 모형(5)는 이러한 성질을 갖도록 대칭성을 일반화시킨 모형이다. 이 모형은 모든 $i\langle j$에 대해

$$P(X=i, Y=j \mid X\langle Y) = P(X=j, Y=i \mid X\rangle Y)$$

를 만족시킨다. 즉, 주대각 위에 나타난 각각의 (i, j)칸의 확률값은 주대각 아래에 나타난 (i, j)칸의 값과 대칭을 이룬다. 이러한 특성때문에 이 모형을 conditional symmetry model (McCullagh, 1978)이라 부른다.

위에서 소개한 5가지 모형들의 적합도를 검정하기 위하여 본 연구에서는 척도편차(scaled deviance; SD)가 점근적으로 카이제곱(χ^2) 통계량을 따른다는 사실을 이용하였다(Agresti, 1990).

Ⅲ. 연구 결과

1. 연구가설의 명확성

표2로부터 얻은 $\prod_o$와 $\prod_e$의 추정값은 $\hat{\prod}_o$=0.611과 $\hat{\prod}_e$=0.381이다. 이 때의 표본 kappa값은 $\hat{K}$=0.373으로 표준오차는 0.057이다. $\hat{K}$의 95% 신뢰구간은 (0.261-0.485)로 주어진 항목에 대한 평가자의 의견일치가 우연에 의한 것만은 아님을 알 수 있다. 이 영역에 있어 관찰된 일치와 우연한 일치 사이의 차이는 최대가능한 차이의 약 37%이다. 평가자들간의 평가가 통계적으로 독립인 경우보다 일치도가 높게 나타났다.

Table 2. Reviewer Ratings on Clearness of Study Hypothesis for Article Check List, with Fitted Values for Quasi-symmetry Model

reviewer A	reviewer B		
	not clear	clear	very clear
not clear	23(23.0)	9(8.4)	2(2.6)
clear	14(14.6)	63(63.0)	31(30.4)
very clear	3(2.4)	16(16.6)	32(32.0)

표3은 표1에서 소개한 5가지 로그선형모형을 적용시킨 결과를 보여준다. Quasi-symmetry모형이 잔차의 SD=0.4, d.f.=1로 가장 잘 적합됨을 알 수 있다. 이때

추정된 모수의 추정치와 표준오차는 표4에 제시되어 있다. 이 모형으로부터 평가자 A와 평가자 B의 도수비(m_{ij}/m_{ji}((i, j)칸의 기대도수/(j, i)칸의 기대도수))를 이용하여 평가자들간의 평가가 불일치했을 경우에 대한 상대적인 평가 경향을 알아볼 수 있다(이영조, 1991). 평가자 A, B의 평가가 일치했을 경우는 이 값이 항상 1을 만족하고, 일치하지 않은 각각의 경우는 다음과 같다. 먼저 범주 (1)명확하지 않다, (2)명확하다의 경우를 보면 m_{12}/m_{21}=1/1.73으로, 평가자 A가 '명확하지 않다'고 평가했을 때 평가자 B가 '명확하다'로 평가한 경우보다, 평가자 A가 '명확하다'라고 했을 때 평가자 B가 '명확하지 않다'고 판단한 경우가 1.73배 많음을 나타낸다. 즉 상대적으로 평가자 A가 좀 더 후한 평가를 한 경향이 있다고 해석할 수 있다.

범주 (2)명확하다, (3)매우 명확하다의 경우를 보면 m_{23}/m_{32}=1.82/1으로, 평가자 A가 '명확하다'고 평가했을 때 평가자 B가 '매우 명확하다'로 평가한 경우가, 평가자 A가 '매우 명확하다'라고 했을 때 평가자 B가 '명확하다'고 판단한 경우보다 1.82배 많음을 의미한다.

이것은 연구가설이 비교적 명확한 논문들에 대해서는 오히려 평가자 B가 평가자 A보다 후하게 평가하였음을 시사한다. 마지막으로 범주 (1)명확하지 않다, (3)매우 명확하다의 경우를 비교해보면 m_{13}/m_{31}=1.05/1로, 두 평가자가 극단적으로 평가한 논문의 개수가 거의 일치함을 알 수 있다. 이들 값을 그림으로 표시하면 아래와 같다.

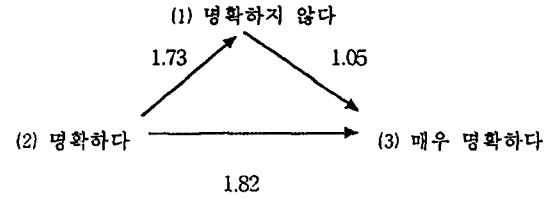


Table 3. Goodness-of-Fit of Models

Fitted model	Scaled Deviance(SD)	d.f.*
Symmetry	6.2	3
Conditional symmetry	5.1	2
Quasi-symmetry	0.4	1
Independence	41.4	4
Quasi-independence	28.8	3

* : degree of freedom

Table 4. Parameter Estimates for Quasi-symmetry Model

Parameter	Estimate	S.E.	P-value
μ	3.47	0.18	0.0001
λ_1^X	-0.14	0.27	0.6104
λ_2^X	0.64	0.18	0.0004
λ_1^Y	-0.19	0.27	0.4634
λ_2^Y	0.04	0.18	0.8475
λ_{12}^{XY}	-1.23	0.25	0.0001
λ_{13}^{XY}	-2.38	0.47	0.0001
λ_{23}^{XY}	-0.69	0.19	0.0002

2) 연구대상의 적절성

표5로부터 얻은 $\prod_o$와 $\prod_e$의 추정값은 $\hat{\prod}_o$=0.819과 $\hat{\prod}_e$=0.741이다. 이때 추정한 표본 카파값 $\hat{K}$=0.299이고, 표준오차는 0.085이다. 관찰된 일치와 우연한 일치 사이의 차이는 최대가능한 차이의 약 30%이다. $\hat{K}$의 95% 신뢰구간은 (0.132-0.466)으로 평가자들간의 평가가 통계적으로 독립인 경우보다는 일치도가 높다는 증거를 보여주었다.

3) 자료수집과정의 신뢰성

표6에서 구한 $\prod_o$와 $\prod_e$의 추정값은 $\hat{\prod}_o$=0.746과

Table 5. Reviewer Ratings on Relevance of Study Population for Article Check List

reviewer A	reviewer B	
	not relevant	relevant
not relevant	11	29
relevant	6	147

$\hat{\prod}_e$=0.577이다. 이때의 표본 카파값 $\hat{K}$=0.399이고, 표준오차는 0.069이다. 관찰된 일치와 우연한 일치 사이의 차이는 최대가능한 차이의 약 40%라 할 수 있고, $\hat{K}$의 신뢰구간은 (0.264-0.534)로 평가자들간의 평가가 통계적으로 독립인 경우보다 일치도가 더 높다는 증거를 보여주었다.

Table 6. Panelist Ratings on Reliability of Data Collection for Article Check List

reviewer A	reviewer B	
	not reliable	reliable
not reliable	34	16
reliable	23	110

4) 결과분석 및 해석

표7에서 얻은 $\prod_o$와 $\prod_e$의 추정값은 $\hat{\prod}_o$=0.539과 $\hat{\prod}_e$=0.244이다. 이때의 표본 카파값 $\hat{K}$=0.390이고, 표준오차는 0.046이다. $\hat{K}$의 95% 신뢰구간은 (0.299-0.481)로 주어진 항목에 대한 평가자의 의견일치가 우연에 의한 것만은 아님을 알 수 있었다. 관찰된 일치와 우연한 일치 사이의 차이는 최대가능한 차이의 약 39%이며, 평가자들간의 평가가 통계적으로 독립적인 경우보다 일치도가 더 높게 나타났다.

표8은 표1에서 소개한 5가지 로그선형모형을 적합시킨 결과를 보여준다. '연구가설의 명확성'에서와 마찬가지로 Quasi-symmetry 모형이 잔차의 SD=3.7, d.f.=3로 가장 잘 적합됨을 알 수 있었다.

Table 8. Goodness-of-Fit of Models

Fitted model	Scaled Deviance(SD)	d.f.
Symmetry	18.8	6
Conditional symmetry	18.5	5
Quasi-symmetry	3.7	3
Independence	157.1	9
Quasi-independence	76.1	8

Table 9. Parameter Estimates for Quasi-symmetry Model

Parameter	Estimate	S.E.	P-value
μ	3.33	0.19	0.0001
λ_1^X	-0.14	0.33	0.6635
λ_2^X	-0.10	0.29	0.7373
λ_3^X	-0.45	0.23	0.0477
λ_1^Y	0.47	0.33	0.1468
λ_2^Y	-0.53	0.29	0.0653
λ_3^Y	0.20	0.23	0.3631
λ_{12}^{XY}	-0.97	0.29	0.0007
λ_{13}^{XY}	-2.46	0.47	0.0001
λ_{14}^{XY}	-26.01	53482.02	0.9996
λ_{23}^{XY}	-0.47	0.27	0.0838
λ_{24}^{XY}	-2.13	0.48	0.0001
λ_{34}^{XY}	-0.49	0.23	0.0355

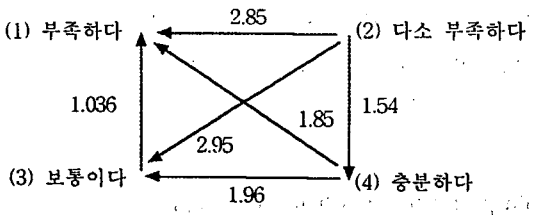
이때 추정된 모수의 추정치와 표준오차는 표9에 제시하였다. 평가자 A와 평가자 B의 상대적인 평가 경향은 다음과 같다. 먼저 범주 (1)부족하다, (2)다소 부족하다의 경우를 보면 m_{12}/m_{21}=1/2.85로, 평가자 A가

Table 7. Reviewer Ratings on Statistical Analysis and Interpretation for Article Check List

reviewer A	reviewer B			
	very poor	poor	average	good
very poor	39(39.0)	7(5.5)	1(2.5)	0(0.0)
poor	14(15.5)	15(15.0)	20(19.4)	4(3.0)
average	4(2.5)	6(22.0)	22(22.0)	10(11.0)
good	0(0.0)	1(2.0)	22(21.0)	28(28.0)

'부족하다'고 평가했을 때 평가자 B가 '다소 부족하다'로 평가한 경우보다, 평가자 A가 '다소 부족하다'라고 했을 때 평가자 B가 '부족하다'고 판단한 경우가 2.85배 많았음을 의미한다. 즉 상대적으로 평가자 A가 좀 더 후한 평가를 했다는 사실을 알 수 있었다.

범주 (2) 다소 부족하다, (3)보통이다의 경우를 살펴보면 m_{23}/m_{32}=2.95/1로, 평가자 A가 '다소 부족하다'고 평가했을 때 평가자 B가 '보통이다'로 평가한 경우가, 평가자 A가 '보통이다'라고 했을 때 평가자 B가 '다소 부족하다'고 판단한 경우보다 2.95배 많음을 의미한다. 즉 상대적으로 평가자 B가 좀 더 후한 평가를 했다는 사실을 나타내었다. 범주 (3) 보통이다, (4)충분하다의 경우를 보면 m_{34}/m_{43}=1/1.92로, 평가자 A가 '보통이다'고 평가했을 때 평가자 B가 '충분하다'로 평가한 경우보다, 평가자 A가 '충분하다'라고 했을 때 평가자 B가 '보통이다'고 판단한 경우가 1.92배 많음을 의미한다. 즉 상대적으로 평가자 A가 좀 더 후한 평가를 했다는 사실을 알려주었다. 위의 사실들을 종합해보면, 범주 (1)부족하다, (2)다소 부족하다와 (3)보통이다, (4)충분하다를 비교했을 때, 각각의 경우에 있어 평가자 A가 좀 더 긍정적인 평가를 한 경향이 있고, 범주 (2)다소 부족하다, (3)보통이다의 경우에 있어서는 반대로 평가자 B가 좀 더 후한 평가를 했다고 할 수 있다. 나머지 범주간의 비교는 다음의 그림으로 제시하였다.



Ⅳ. 고 찰

본 연구에서는 복수 평가자들간의 평가가 어느 정도 일치하는지를 검정하는 통계적인 방법을 소개하고, 그러한 방법들을 직접 적용한 결과를 보고하고자 하였다. 카파 통계량과 로그선형모형들(즉, Symmetry model, Conditional symmetry model, Quasi-symmetry model, Independence model, Quasi-independence model)은 원래 두 명의 평가자에게 적용되는 방법으로 소개되어 있지만, 여러 명의 평가자들 중 절대적인 신뢰가 가는 평가자가 있다고 가정할 수 있는 경우에는 이들 통계적 방법을 본 연구에서와 같이 적용하여 여러 평가자들간의 신뢰도를 평가할 수 있을 것이다.

카파값은 평가자들간의 평가가 서로 일치하고 있는지, 즉 신뢰도가 어느 정도인지를 알려준다. 또한 로그선형모형들을 적합시켜봄으로써 평가자들간의 평가가 불일치하는 경우에 있어 평가자들간의 평가 경향까지도 설명할 수가 있다. 이러한 방법들은 방사선과에서 과거 수 년동안 여러 병원에서 촬영하여 보관되어 있는 필름들을 수 명의 방사선과 전문의들이 공동으로 재평가하여 표준화된 진단명을 내리고자 하는 경우나, 여러 의료기관에서 공동으로 연구를 수행하는 경우에 각 의료기관에서 조직학적 진단을 위하여 제작한 슬라이드표본을 중앙검사센터에서 취합하여 여러 병리학자들이 공동으로 진단을 내리는 경우, 또는 많은 수의 연구논문을 여러 명의 연구자들이 공동으로 평가하는 연구 등과 같이 연구대상의 규모가 큰 역학연구를 수행하는 경우에, 연구에 참여한 공동연구자들간의 평가기준의 신뢰도를 평가하고 서로간의 평가결과를 조절하는데 도움을 줄 수 있으리라 판단된다.

로그선형모형은 분류변수가 명목형인지 순위형인지에 따라 다양하게 적용될 수 있다. 본 논문에서 소개한 모형들은 분류변수가 명목형과 순위형 모두에 사용 가능한 모형들이다. 순위형 분류변수로 이루어진 분할표를 분석할 경우에는 '순위'의 영향을 고려함으로써 본 논문에서 소개한 모형들보다 더욱 parsimonious한 모형의 적합을 고려할 수도 있다. 예를 들어, Quasi-association모형은 순위형 분류변수에서만 분석 가능한 모형이다. 본 연구에서는 이 모형의 적합도가 통계적으로 유의하지 않아 소개하지 않았다(Agresti, 1990).

본 연구 사례에서 7명의 논문 평가자들간의 신뢰도는 '연구가설의 명확성'을 평가하는 항목에서 카파값이 0.373, '연구대상의 적절성'을 평가하는 항목에선 0.299, '자료수집과정의 신뢰성'을 평가하는 항목에서 0.399, '결과분석 및 해석' 부분을 평가하는 항목에서 0.390으로 나타났으며, 모두 통계적으로 유의한 일치도를 보였다. Landis 등(1977)은 카파값의 크기에 따라 0보다 작을 때는 불량(poor), 0에서 0.20사이는 약간의(slight), 0.21에서 0.4 사이는 상당한(fair), 0.41에서 0.60사이는 중등도의(moderate), 0.61부터 0.80사이는 실질적인(substantial), 0.81에서 1.00사이는 거의 완벽한(almost perfect) 일치도를 가지는 것으로 평가하는 기준을 제안한 바 있다. 이러한 분류기준에 따르면 본 연구에서 나타난 평가자들간의 일치도는 상당한 수준의 일치도를 나타내고 있는 것으로 평가할 수 있겠다. 그러나 이러한 기준은 절대적인 것으로 받아들이기 어려운 것으로 알려져 있다. 즉, 연구대상에 따라서 다른 의미를 가지는데, 물리 화학등의 자연과학 분야나 공학분야에서는 1에 아주 가까운 값이어야 의미를 가지는 것으로 받아들여지는 반면에, 인문사회과학 분야에서는 0.1이나 0.2정도도 유의한 의미를 가지는 것으로 평가될 수 있기 때문이다. 본 연구도 여러 가지 요인이 복합적으로 관여하는 사회과학문야 연구의 특성을 가지므로 유의한 의미를 가지는 결과로 받아들여질 수 있을 것으로 판단된다. 이들 카파값은 평가자들간의 관찰된 일치와 우연에 의한 일치간의 차이가 최대 가능한 차이와 비교해 보았을 때 각각 37.3%, 29.9%, 39.9%, 및 39.0% 임을 의미한다. 항목별로는 '자료수집과정의 신뢰성'과 '결과분석 및 해석'을 평가하는데 있어 평가자들간의 평가가 가장 높은 일치도를 보여 주었다. 카파값 자체로는 '연구대상의 적절성'이 상대적으로 가장 낮은 일치도를 보였다.

켄달의 타우-b값을 비교해 보면, 각각 0.47, 0.34, 0.40, 0.67로 유의확률이 0.01%로 매우 유의하여 평가자들간의 유의한 관련성이 있음을 보여주었다. 특히 '결과분석 및 해석'부분에 있어서는 관련성이 매우 높음을 알 수 있었다.

로그선형모형을 이용한 분석에서 다섯 가지 모형가운데, Quasi-symmetry 모형이 연구가설의 명확성 평가, 자료수집과정의 신뢰성 평가, 및 결과분석 및 해석 부분의 평가에서 공히 가장 높은 적합도를 나타내었다. 각 항목별로 살펴보면 연구가설의 명확성 평가와 결과분석 및 해석부분의 평가에 있어서 평가자 A는 다른 평가자들에 비하여 부족하다는 평가에서는 다른 평가자들에 비하여 다소 후하게 평가하여 극단적으로 부정적인 평가를 하지 않는 경향을 나타내었다.

본 연구결과는 향후 여러 명의 연구자들이 다수 검사자료를 대상으로 공동으로 진단을 내리거나, 대규모 연구결과를 복수의 평가자들이 평가하는 임상의학 연구를 수행하는 경우에 연구자간 신뢰도를 평가하고 연구자들간의 평가 경향을 파악하는데 매우 유용하게 활용될 수 있을 것으로 판단된다.

참고 문헌

김병수, 안윤기, 윤기중, 윤상운. SPSS를 이용한 통계자료분석,박영사, 1987.

박병주. 서울대학교병원에서의 임상의학연구 현황. 임상시험활성화를 위한 워크샵. 서울대학교병원 임상의학연구소, 1994, 쪽 3-17.

이영조, 배상수, 한달선. Quasi-Symmetric모형을 이용한 의료기관 유형간의 환자이동 분석, 한국보건통계학회지 1991;16:1-9.

허명회. SAS범주형 데이타 분석, 자유아카데미, 1995.

Agresti A. Categorical Data Analysis. New York: John Wiley & Sons, 1990.

Caussinus H. Contribution a l'analyse statstique des tableaux de correlation. Ann. Fac. Sci. Univ. Toulouse 1965;29:77-182.

Cohen J. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 1960;20:37-46.

Fleiss JL, Cohen J, Everitt BS. Large-sample standard errors of kappa and weighted kappa. Psychol. Bull. 1969;72:323-327.

Kendall MG. A new measure of rank correlation. *Biometrika* 1938;30:81-93.

Kruskal WH. Ordinal measures of association. Journal of American Statistical Association.

1958;53:814-861.
Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.
McCullagh P. A class of parametric models for the analysis of square contingency tables with ordered categories. Biometrika 1978;65:413-418.

〈부 록〉

Article Check List

1. Idno – □□□□□□
2. Reviewer – □[1] Kwon □[2] Kim □[3] Bae □[4] Shin □[5] Ha □[6] Han □[7] PI
3. Natural – □[1] Survey □[2] Etiology □[3] Diagnos □[4] Treatment □[5] Prognosis □[6] Prevention □[7] Undefined
4. Design – □[1] Case report □[2] Case series □[3] Cross-sectional □[4] Case-control
   □[5] Cohort □[6] Clinical trial □[7] Animal Lab □[8] Review/Meta □[9] Undefined
5. Discase – □[1] Inf □[2] Ca □[3] Endo/Nut/Meta/Imm □[4] Blood □[5] Mental □[6] Neuro
   □[7] CardioV □[8] Resp □[9] GI □[10] GU □[11] Preg □[12] Skin □[13] MS
   □[14] Congen □[15] Perinatal □[16] Sx/Sign/Undefined □[17] Injury/Poisoning □[18] Otherwise

1. 연구 가설은 제시되어 있는가?	□N □Y □NA □DK
2. 있다면 그 가설은 구체적인가?	□N □1 □2 □Y □NA □DK
3. 연구가설에 대해 연구설계는 적절한가?	□N □1 □2 □Y □NA □DK
4. 연구대상의 선정기준은 제시되고 있는가?	□N □1 □2 □Y □NA □DK
5. 연구대상의 기본적 특성은 제시되어 있는가?	□N □Y □NA □DK
6. 자료수집과정에 대한 기술은 있는가?	□N □Y □NA □DK
7. 있다면 연구과정의 재현성이 보장될 정도로 자세한가?	□N □1 □2 □Y □NA □DK
8. 결과분석에 있어 사용된 통계방법은 적절한가?	□N □1 □2 □Y □NA □DK
9. 관련도지표의 통계적 안정성은 어떻게 제시되고 있는가?	□P □CI □not □NA
10. 통계분석 결과에 대한 해석은 타당한가?	□N □1 □2 □Y □NA □DK
11. 결과해석이 연구목적에 부합하는가?	□N □1 □2 □Y □NA □DK