Hanyang Medical Reviews

Hanyang Med Rev 2015;35:44-49
http://dx.doi.org/10.7599/hmr.2015.35.1.44
pISSN 1738-429X eISSN 2234-4446

체계적 문헌고찰에서 평가자 간의 신뢰도 측정

박창언[1] · 김현정[2]

[1]제주대학교 의학전문대학원, [2]제주대학교 의학전문대학원 마취통증의학과

Measurement of Inter-Rater Reliability in Systematic Review

Chang Un Park[1], Hyun Jung Kim[2]

[1]Jeju National University School of Medicine, Jeju; [2]Department of Anesthesiology and Pain Medicine, Jeju National University School of Medicine, Jeju, Korea

Inter-rater reliability refers to the degree of agreement when a measurement is repeated under identical conditions by different raters. In systematic review, it can be used to evaluate agreement between authors in the process of extracting data. While there have been a variety of methods to measure inter-rater reliability, percent agreement and Cohen's kappa are commonly used in the categorical data. Percent agreement is an amount of actually observed agreement. While the calculation is simple, it has a limitation in that the effect of chance in achieving agreement between raters is not accounted for. Cohen's kappa is a more robust method than percent agreement since it is an adjusted agreement considering the effect of chance. The interpretation of kappa can be misled, because it is sensitive to the distribution of data. Therefore, it is desirable to present both values of percent agreement and kappa in the review. If the value of kappa is too low in spite of high observed agreement, alternative statistics can be pursued.

Key Words: Agreement; Inter-Rater; Kappa; Rater; Reliability

Correspondence to: Hyun Jung Kim
우690-767, 제주특별자치도 제주시 아란13길 15, 제주대학교 의학전문대학원 마취통증의학과
Department of Anesthesiology and Pain Medicine, Jeju National University School of Medicine, 15 Aran 13gil, Jeju 690-767, Korea
Tel: +82-64-717-2029
Fax: +82-64-717-2042
E-mail: hjanesthesia@empas.com

Received 27 November 2014
Revised 12 December 2014
Accepted 26 December 2014

서 론

체계적 문헌고찰(systematic review)에서 데이터 추출은 연구 목적에 맞는 정보를 검색된 일차 문헌에서 뽑아내는 과정으로 연구의 질을 결정하는 중요한 과정 중에 하나이다. 한 명의 연구자가 데이터를 추출할 경우 평균 8%의 유의한 데이터가 누락된다고 하며, 두 명의 연구자가 각각 독립적으로 데이터를 추출할 경우 유의한 데이터의 누락 없이 거의 완벽하게 데이터를 추출할 수 있다고 한다[1]. 따라서 데이터 추출 시에는 최소한 두 명의 연구자가 각각 독립적으로 데이터를 추출한 후 그 일치 여부를 확인하는 과정이 필요한데 이때 사용되는 지표가 평가자 간의 신뢰도(inter-rater reliability, IRR)이다[2].

신뢰도(reliability)란 측정한 값들이 얼마나 일관적이고 동일한 결과를 보이는가 하는 개념이다. IRR은 여러 명의 평가자(rater)들이 동일한 조건에서 어떤 대상에 대해 평가를 할 때 의견이 일치되는 정도를 나타내는 척도로써 높은 IRR은 평가자료나 평가방법이 안정되고 일관성이 있으며 정확하다는 것을 의미한다[3]. IRR을 측정하는 방법은 자료의 종류와 평가자의 수에 따라 여러 가지가 있지만 본 종설에서는 두 명의 평가자가 범주형 자료(categorical data)를 평가하는 경우로 한정하여 대표적인 몇 가지 방법에 대해 살펴보도록 하겠다.

본 론

1. 퍼센트 일치도(percent agreement)

일치도(agreement)란 한 표본을 여러 번 반복 측정한 결과가 서로 어느 정도 일치하는가를 나타내는 신뢰도 평가의 방법 중 하나이다. 두 명의 평가자 A와 B가 각각 독립적으로 N개의 자료를 '1' 또

는 '2'로 분류하는 경우 Table 1 (A)와 같이 정리할 수 있다. 이때 *a*와 *d*는 두 명의 평가자가 같은 범주로 분류한 자료 수이고 *b*와 *c*는 다른 범주로 분류한 자료 수가 된다.

퍼센트 일치도는 전체 자료 중에 두 명의 평가자가 같은 범주로 분류한 자료의 퍼센트이다. Table 1 (A)에서 퍼센트 일치도는 아래와 같이 정의된다.

$$\text{Percent agreement (\%)} = \frac{a+d}{N} \times 100$$

Table 1 (B)와 같은 자료가 주어지면 퍼센트 일치도는 아래와 같다.

$$\text{Percent agreement (\%)} = \frac{40+30}{100} \times 100 = 70\%$$

퍼센트 일치도는 계산이 간단하고 평가자 수에 관계없이 적용할 수 있다는 장점이 있다. 그러나 평가자들이 우연히 자료를 동일한 범주로 분류할 확률을 포함하고 있기 때문에 실제로 평가자들의 의견이 일치된 비율보다 높은 값을 보이는 것이 단점이다[4].

2. Cohen의 카파통계량(Cohen's kappa statistic)

1) Cohen의 카파통계량의 정의와 해석

1960년에 Cohen이 제시한 카파(Cohen's kappa, κ)통계량은 IRR을 측정하는 방법으로 가장 널리 사용되고 있는 통계량이다[5]. 카파통계량에서는 평가자들이 우연히 자료를 동일한 범주로 분류할 확률을 보정한 일치도를 사용한다.

$$\kappa = \frac{Po - Pc}{1 - Pc}$$

Po = the proportion of units for which agreement is actually observed

Pc = the proportion of units for which agreement is expected by chance

Table 1. Data in 2×2 table format

		Rater B		Total
		1	2	
Rater A	1	*a*	*b*	A_1
	2	*c*	*d*	A_2
	Total	B_1	B_2	*N*

(A) Distribution of *N* subjects by 2 raters and 2 scales.

		Rater B		Total
		1	2	
Rater A	1	40	10	50
	2	20	30	50
	Total	60	40	100

(B) Example data of 2×2 table format.

여기서 Po는 평가자들이 같은 범주로 분류한 자료의 비율이고, Pc는 평가자들이 우연히 자료를 동일한 범주로 분류할 기대비율을 의미한다. 카파통계량은 관찰된 일치 비율에서 우연에 의한 일치 비율을 뺀 값과 평가자들의 평가가 완벽하게 일치할 비율인 '1'에서 우연에 의한 일치 비율을 뺀 값의 비로 정의된다. 여기서 평가자들은 독립적으로 평가를 시행해야 하며, 평가대상도 독립적이고 서로 겹치는 구간이 없는 명목척도(nominal scale)여야 한다.

Table 1 (A)와 (B)에서 Po는 아래와 같다.

$$Po = \frac{a+d}{N} = \frac{40+30}{100} = 0.7$$

Table 1 (A)와 (B)에서 Pc는 평가자 A와 B가 각각 평가대상을 우연히 '1'로 분류할 확률과 '2'로 분류할 확률을 합한 값으로 정의된다.

$$Pc = \frac{A_1}{N} \times \frac{B_1}{N} + \frac{A_2}{N} \times \frac{B_2}{N} = \frac{50}{100} \times \frac{60}{100} + \frac{50}{100} \times \frac{40}{100} = 0.5$$

따라서 κ는 아래와 같이 얻어진다.

$$\kappa = \frac{Po - Pc}{1 - Pc} = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

위의 경우를 확대하여 Table 2처럼 두 명의 평가자가 *N*개의 자료를 *q*개의 범주로 분류할 경우 κ는 아래와 같다.

$$\kappa = \frac{Po - Pc}{1 - Pc}$$

$$Po = \frac{1}{N} \sum_{i=1}^{q} n_{ii}$$

$$Pc = \frac{1}{N^2} \sum_{i=1}^{q} A_i B_i$$

다른 통계량과 마찬가지로 카파통계량도 κ값만을 제시하는 것보다 신뢰구간(confidence intervals)을 같이 제시하는 것이 통계적 추론에 도움이 된다. 카파통계량의 표준오차(standard error of kappa, $\delta\kappa$)와 신뢰구간은 아래와 같이 구할 수 있다.

$$\delta\kappa = \sqrt{\frac{Po\,(1 - Po)}{N\,(1 - Pc)^2}}$$

95% Confidence intervals = $\kappa \pm 1.96 \times \delta\kappa$

99% Confidence intervals = $\kappa \pm 2.58 \times \delta\kappa$

Table 2. Data for calculating of a kappa statistic

		Rater B				Total
		1	2	-	*q*	
Rater A	1	n_{11}	n_{12}	-	n_{1q}	A_1
	2	n_{21}	n_{22}	-	n_{2q}	A_2
	-			-		-
	q	n_{q1}	n_{q2}	-	n_{qq}	A_q
	Total	B_1	B_2	-	B_q	*N*

카파통계량은 -1부터 +1의 값을 가질 수 있는데, 1은 평가자 간에 의견이 모두 실제로 일치하는 것을 의미하고 0은 완벽히 우연에 의한 일치를 의미한다. 음의 값은 평가자 간에 반대되는 의견을 의미하지만 실제로 음의 값이 나오는 연구는 드물다.

카파통계량은 연구에 따라 다양하게 해석될 수 있다. 그중 Landis와 Koch가 제안한 분류가 널리 사용되고 있는데 κ값을 6단계로 나누어 일치 수준을 평가한다[6]. 즉, 카파값에 따라 ≤0 부족한(poor), 0.0-0.20 약간(slight), 0.21-0.40 조금 큰(fair), 0.41-0.60 중간의(moderate), 0.61-0.80 상당한(substantial), 0.81-1.0 거의 완벽한(almost perfect) 일치를 의미한다.

2) SPSS®에서 카파통계량 구하기

널리 사용되는 통계 프로그램인 SPSS®을 이용하여도 카파통계량을 구할 수 있다[7]. Table 1 (B)를 이용하여 카파통계량을 구하기 위해서는 데이터 창을 열고 변수로 'Rater A', 'Rater B', 'Total'을 입력한다. 평가자 A와 평가자 B가 평가대상을 '1'로 분류한 경우는 1로 입력하고 '2'로 분류한 경우는 2로 입력한 후 각 항목에 해당하는 평가대상 수를 Fig. 1과 같이 입력한다.

평가대상 수에 가중치를 부여하기 위하여 [데이터] → [가중 케이스] → [가중 케이스 지정]으로 들어가 빈도변수에 'Total'을 선택한다. 다음으로 [분석] → [기술통계량] → [교차분석]으로 들어가 행에 'Rater A'를 선택하고 열에 'Rater B'를 선택한다. 마지막으로 통계량으로 들어가 [카파(κ)]를 선택한 후 계산을 시행하면 Fig. 2와 같은 결과가 나온다.

분석 결과 $\kappa = 0.4$, $\delta\kappa = 0.09$, 유의확률 $P < 0.001$임을 알 수 있다. 신뢰구간은 결과에는 표시되어 있지 않지만 아래와 같이 구할 수 있다[5].

$$95\%\ \text{Confidence intervals} = \kappa \pm 1.96 \times \delta\kappa = 0.4 \pm 1.96 \times 0.09 = 0.224 \text{ to } 0.576$$

$$99\%\ \text{Confidence intervals} = \kappa \pm 2.58 \times \delta\kappa = 0.4 \pm 2.58 \times 0.09 = 0.168 \text{ to } 0.632$$

3) 카파통계량의 한계점

카파통계량은 평가자 간의 신뢰도를 측정할 때 널리 사용되고 있는 방법이지만, 관찰된 일치 비율 *Po*가 동일하더라도 자료의 분포에 따라 κ가 크게 변하는 한계점이 있다. Feinstein와 Cicchetti는 그 원인을 크게 두 가지로 분석하였다[8,9].

첫째, 각 평가자가 평가대상들을 두 범주에 균일하게 할당한 경우, 즉 Table 1에서 $\frac{A_1}{N}$과 $\frac{B_1}{N}$의 값이 0.5에 가까울수록 주변분포가 균형적(balanced marginal distribution)이라고 하는데 관찰된 일치비율 *Po*가 동일하더라도 κ는 주변분포가 균형적일 때 그렇지 않은 경우보다 큰 값을 보인다[10]. 예를 들어, Table 3에서 (A)와 (B)의

RaterA * RaterB 교차표

빈도

		RaterB		
		1	2	전체
RaterA	1	40	10	50
	2	20	30	50
전체		60	40	100

대칭적 측도

		값	점근 표준오차[a]	근사 T 값[b]	근사 유의확률
일치 측도	카파	.400	.090	4.082	.000
유효 케이스 수		100			

a. 영가설을 가정하지 않음.

b. 영가설을 가정하는 점근 표준오차 사용

Fig. 2. Display of SPSS® results for the kappa test.

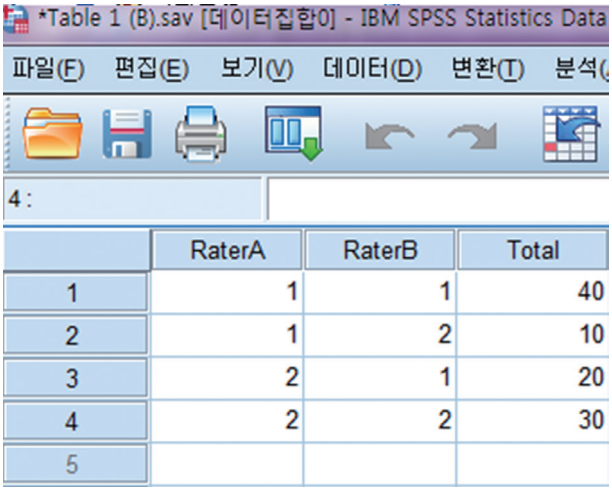

Fig. 1. SPSS® data file in count form.

Table 3. Example data of 2×2 table format showing data distribution with or without balanced marginal distribution

		Rater B		Total
		1	2	
Rater A	1	40	10	50
	2	10	40	50
	Total	50	50	100

$Po = \frac{40+40}{100} = 0.8$

$Pc = \frac{50}{100} \times \frac{50}{100} + \frac{50}{100} \times \frac{50}{100} = 0.5$

$\kappa = \frac{0.8-0.5}{1-0.5} = 0.6$

(A) Data distribution with balanced marginal distribution.

		Rater B		Total
		1	2	
Rater A	1	75	10	85
	2	10	5	15
	Total	85	15	100

$Po = \frac{75+5}{100} = 0.8$

$Pc = \frac{85}{100} \times \frac{85}{100} + \frac{15}{100} \times \frac{15}{100} = 0.745$

$\kappa = \frac{0.8-0.745}{1-0.745} = 0.216$

(B) Data distribution with unbalanced marginal distribution.

Table 4. Example data of 2×2 table format showing data distribution with or without marginal homogeneity

		Rater B		Total
		1	2	
Rater A	1	40	20	60
	2	20	20	40
	Total	60	40	100

$Po = \frac{40+40}{100} = 0.6$

$Pc = \frac{60}{100} \times \frac{60}{100} + \frac{40}{100} \times \frac{40}{100} = 0.52$

$\kappa = \frac{0.6 - 0.52}{1 - 0.52} = 0.167$

(A) Data distribution with marginal homogeneity.

		Rater B		Total
		1	2	
Rater A	1	30	30	60
	2	10	30	40
	Total	40	60	100

$Po = \frac{30+30}{100} = 0.6$

$Pc = \frac{60}{100} \times \frac{40}{100} + \frac{40}{100} \times \frac{60}{100} = 0.48$

$\kappa = \frac{0.6 - 0.48}{1 - 0.48} = 0.231$

(B) Data distribution without marginal homogeneity.

관찰된 일치 비율 *Po*는 0.8로 동일하지만, κ는 각각 0.6과 0.216으로 주변분포가 균형적인 (A)에서 κ가 더 높게 나타남을 알 수 있다.

둘째, Table 1에서 $A_1 = B_1$인 경우 이를 주변동질성(marginal homogeneity)을 만족한다고 하는데 관찰된 일치 비율 *Po*가 동일하더라도 κ는 주변동질성을 만족할 때 그렇지 않은 경우보다 작은 값을 보인다[10]. 예를 들어, Table 4에서 (A)와 (B)의 관찰된 일치 비율 *Po*는 0.6으로 동일하지만, κ는 각각 0.167과 0.231로 A_1과 B_1의 값이 동일한 (A)에서 κ가 더 낮게 나타남을 알 수 있다.

위와 같이 자료의 분포에 따라 관찰된 일치 비율 *Po*가 비슷하더라도 κ가 다르게 측정되거나 *Po*가 높더라도 κ가 낮게 측정되는 특성이 카파통계량의 단점이라고 할 수 있다.

3. 기타 통계량

위에서 언급한 바와 같이 카파통계량은 자료의 분포에 따라 κ가 크게 변하는 단점이 있기 때문에 이와 같은 단점을 보완하는 다른 여러 가지 통계방법이 제시되고 있다.

1) 양의 일치도(positive agreement, P_{pos})와 음의 일치도(negative agreement, P_{neg})

Cicchetti와 Feinstein는 κ를 제시할 때 P_{pos}와 P_{neg}를 같이 제시할 것을 주장하였다[9]. Table 1 (A)에서 P_{pos}는 두 평가자가 자료를 '1'로 분류할 평균적인 비율이고 P_{neg}는 자료를 '2'로 분류할 평균적인 비율로 아래와 같이 정의된다.

$$P_{pos} = \frac{a}{\left(\frac{A_1 + B_1}{2}\right)} = \frac{2a}{A_1 + B_1}$$

$$P_{neg} = \frac{d}{\left(\frac{A_2 + B_2}{2}\right)} = \frac{2d}{A_2 + B_2}$$

Table 3 (A)와 (B)에서 P_{pos}와 P_{neg}는 아래와 같다.

Table 3 (A)

$$P_{pos} = \frac{2a}{A_1 + B_1} = \frac{2 \times 40}{50 + 50} = 0.8$$

$$P_{neg} = \frac{2d}{A_2 + B_2} = \frac{2 \times 40}{50 + 50} = 0.8$$

$$P_{pos} - P_{neg} = 0$$

Table 3 (B)

$$P_{pos} = \frac{2a}{A_1 + B_1} = \frac{2 \times 75}{85 + 85} = 0.882$$

$$P_{neg} = \frac{2d}{A_2 + B_2} = \frac{2 \times 5}{15 + 15} = 0.333$$

$$P_{pos} - P_{neg} = 0.549$$

P_{pos}와 P_{neg}의 차이가 크면, 즉 $P_{pos} - P_{neg}$의 값이 크면 자료의 분포가 κ에 미치는 영향이 큰 것을 의미한다. Table 3 (A)와 (B)의 $P_{pos} - P_{neg}$의 값은 각각 0과 0.549로 (B)의 κ값은 (A)보다 자료의 분포에 영향을 많이 받았음을 알 수 있다.

2) 자유 주변분포를 가정한 카파(kappa for free marginal, κ_n)

자유 주변분포를 가정한 카파에서는 평가자가 평가대상을 q개로 분류할 때 각각의 항목으로 분류할 확률이 동일하다고 가정하여 우연에 의한 일치 비율 Pc_n를 구한다($Pc_n = \frac{1}{q}$)[11]. Table 3 (B)에서 $\kappa_n = 0.6$으로 $\kappa = 0.216$보다 높은 값을 가진다.

$$\kappa_n = \frac{Po - Pc_n}{1 - Pc_n} = \frac{Po - \frac{1}{q}}{1 - \frac{1}{q}}$$

Table 3 (B)

$$\kappa_n = \frac{Po - Pc_n}{1 - Pc_n} = \frac{Po - \frac{1}{q}}{1 - \frac{1}{q}} = \frac{0.8 - \frac{1}{2}}{1 - \frac{1}{2}} = 0.6$$

3) 유별률과 비뚤림을 보정한 카파(prevalence–adjusted and bias–adjusted kappa, PABAK)

PABAK는 유별률(prevalence)과 비뚤림(bias)을 보정한 것으로 관찰된 일치 비율 *Po*와 선형관계를 가진다[12]. Table 3 (B)에서 PABAK= 0.6으로 $\kappa = 0.216$보다 높은 값을 가진다.

PABAK= $2Po - 1$

Table 3 (B) PABAK = $2Po - 1 = 2 \times 0.8 - 1 = 0.6$

4) AC_1 통계량(AC_1 statistic)

AC_1 통계량에서는 우연에 의한 일치 비율 Pc_γ를 Table 1에서 아래와 같이 정의한다[13]. Table 3 (B)에서 $AC_1 = 0.732$로 $\kappa = 0.216$보다 높은 값을 가진다.

$$AC_1 = \frac{Po - Pc_\gamma}{1 - Pc_\gamma}$$

$$Pc_\gamma = 2P_+ (1 - P_+)$$

$$P_+ = \frac{1}{N}\left(\frac{A_1 + B_1}{2}\right)$$

Table 3 (B)

$$P_+ = \frac{1}{N}\left(\frac{A_1 + B_1}{2}\right) = \frac{1}{100}\left(\frac{85 + 85}{2}\right) = 0.85$$

$$Pc_\gamma = 2P_+ (1 - P_+) = 2 \times 0.85 (1 - 0.85) = 0.255$$

$$AC_1 = \frac{Po - Pc_\gamma}{1 - Pc_\gamma} = \frac{0.8 - 0.255}{1 - 0.255} = 0.732$$

4. 체계적 고찰에서 IRR 측정

체계적 고찰에서는 데이터를 누락 없이 완벽하게 추출하기 위하여 최소한 두 명 이상의 연구자가 각각 독립적으로 데이터를 추출한 후 추출된 데이터에 대한 평가가 연구자들 사이에 얼마나 일치하는지 확인해야 한다[1]. 연구자들 사이에 의견 불일치는 대부분 토론을 통해 쉽게 일치를 이룰 수 있다. 의견 불일치의 원인은 연구 주제에 대한 이해 부족, 연구 방법이나 데이터 추출 과정에 대한 이해 부족, 데이터 추출 과정에서의 실수 등 토론과 교육을 통해 교정될 수 있는 경우가 가장 많다. 그러나 데이터에 대한 해석 자체가 다르다면 해당 자료의 원저자에게 의견을 구하거나 제3자의 도움을 받아 의견일치를 이루어야 한다. 이러한 모든 과정은 기록으로 남겨 두어 누구나 검토가 가능하도록 해야 하며, 마지막까지 의견이 일치되지 않은 데이터가 있다면 논문에서 기술을 해야 한다[2].

연구자들 사이에 의견이 일치되는 정도는 IRR을 측정해서 확인할 수 있다. 만약 IRR이 낮다면 데이터 추출 과정에 대한 재검토를 할 필요성이 있음을 의미한다. IRR 측정은 데이터를 추출하는 과정 중 데이터를 '선택' 또는 '평가'하는 여러 단계에서 각각 시행되어야 하지만, 논문에서 모든 과정에 대해 기술하는 것이 현실적으로 어렵기 때문에 그중에서 가장 중요한 단계의 IRR만을 기술하는 것이 일반적이다. 체계적 고찰에서 IRR이 어느 정도 수준이 되어야 하는가는 아직 명확하게 합의되지 않았다. 일반적으로 많은 연구에서 퍼센트 일치도는 80%, 카파통계량은 0.6 이하인 경우 그 자료는 적절치 않은 것으로 간주하고 있다[14].

예를 들어, Stoller 등[15]은 체계적 고찰에서 다음과 같이 IRR에 대해 기술하였다.

Methods

Analysis

(중략) *Cohen's kappa were calculated and interpreted in accordance with Landis and Koch's benchmarks for assessing the agreement between raters: poor (≤ 0), slight (0.0 to 0.20), fair (0.21 to 0.40), moderate (0.41 to 0.60), substantial (0.61 to 0.80), and almost perfect (0.81 to 1.0).*

Results

Risk of bias within studies

The methodological quality of the studies included is summarised in Table 2. The most common quality problems were absence of blinding of the assessors (8 studies), absence of concealment of allocation (7 studies), and absence of an intention-to-treat analysis (4 studies). The inter-rater agreement of the quality assessment was considered to be very good (Kappa: 0.81, SE: 0.06, CI95%: 0.70-0.92).

위의 연구에서는 메타분석에 포함된 연구들을 대상으로 방법론적인 질적평가를 시행하였는데, 그에 대한 IRR을 측정하기 위해 카파통계량을 이용하였다. 그 결과 $\kappa = 0.81$, 표준오차 0.06, 95% 신뢰구간 0.70-0.92로 IRR은 거의 완벽하게 일치하는 수준으로 나타났다.

결 론

체계적 고찰에서는 최소한 두 명 이상의 연구자가 각각 독립적으로 데이터를 추출한 후 추출된 데이터에 대한 평가가 연구자들 사이에 얼마나 일치하는지 확인하기 위해 IRR을 측정하는데, 이때 가장 많이 사용되는 통계방법은 퍼센트 일치도와 Cohen의 카파통계량이다. 퍼센트 일치도는 전체 자료 중에 평가자들이 같은 범주로 분류한 자료의 비율로 계산이 간단하고 평가자 수에 관계없이 적용할 수 있지만 평가자들이 우연히 자료를 동일한 범주로 분류할 확률을 포함하고 있기 때문에 실제 일치된 비율보다 높은 값을 보이게 된다. Cohen의 카파통계량은 관찰된 일치 비율에서 평가자들이 우연히 자료를 동일한 범주로 분류할 확률을 보정한 일치도를 사용하는데 자료의 분포에 따라 κ값이 크게 달라지는 단점이 있기 때문에 해석에 주의가 필요하다. 일반적으로 IRR을 제시할 때에는 퍼센트 일치도와 Cohen의 카파통계량을 같이 표시하는 것이 바람직하며, 만약 관찰된 일치 비율이 높음에도 불구하고 κ값이 낮을 경우에는 카파통계량의 단점을 보완해주는 다른 통계방법을 사용해 볼 수 있다.

REFERENCES

1. Edwards P, Clarke M, DiGuiseppi C, Pratap S, Roberts I, Wentz R. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. Stat Med 2002;21:1635-40.
2. Higgins JP, Deeks JJ. Cochrane handbook for systematic reviews of interventions: chapter 7. Selecting Studies and Collecting Data. New York. John Wiley & Sons Ltd 2008:151-6.
3. Viera AJ, Garrett JM. Understanding interobsever agreement: the kappa statistic. Fam Med 2005;37:360-3.
4. Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability:

key concepts, approaches, and applications. Res Social Adm Pharm 2013;9:330-8.
5. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37-46.
6. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
7. SPSS statistics base 17.0 user's guide. Chicago: SPSS Inc.:289-93.
8. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543-9.
9. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43:551-8.
10. Kim MS, Song KJ, Nam CM, Jung I. A study on comparison of generalized kappa statistics in agreement analysis. KJAS 2012;25:719-31.
11. Rrennan RL, Prediger DJ. Coefficient kappa: some uses, misuses, and alternatives. Educ Psychol Meas 1981;41:687-99.
12. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46:423-9.
13. Gwet KL. Handbook of inter-rater reliability. 3rd ed. Gaithersburg: Advanced analytics; 2012:15-28.
14. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22:276-82.
15. Stoller O, de Bruin ED, Knols RH, Hunt KJ. Effects of cardiovascular exercise early after stroke: systematic review and meta-analysis. BMC Neurol 2012;12:45.