EMJ
THE EWHA MEDICAL JOURNAL

Review Article

Ewha Med J 2017;40(1):9-16
https://doi.org/10.12771/emj.2017.40.1.9
pISSN 2234-3180 • eISSN 2234-2591

검사법 평가: 검사법 비교와 신뢰도 평가

공경애
이대목동병원 임상시험센터

Statistical Methods: Reliability Assessment and Method Comparison

Kyoung Ae Kong
Clinical Trial Center, Ewha Womans University Mokdong Hospital, Seoul, Korea

The reliability of clinical measurements is critical to medical research and clinical practice. Newly proposed methods are assessed in terms of their reliability, which includes their repeatability, intra- and interobserver reproducibility. In general, new methods that provide repeatable and reproducible results are compared with established methods used clinically. This paper describes common statistical methods for assessing reliability and agreement between methods, including the intraclass correlation coefficient, coefficient of variation, Bland-Altman plot, limits of agreement, percent agreement, and the kappa statistic. These methods are more appropriate for estimating reliability than hypothesis testing or simple correlation methods. However, some methods of reliability, especially unscaled ones, do not clearly define the acceptable level of error in real size and unit. The Bland-Altman plot is more useful for method comparison studies as it assesses the relationship between the differences and the magnitude of paired measurements, bias (as mean difference), and degree of agreement (as limits of agreement) between two methods or conditions (e.g., observers). Caution should be used when handling heteroscedasticity of difference between two measurements, employing the means of repeated measurements by method in methods comparison studies, and comparing reliability between different studies. Additionally, independence in the measuring processes, the combined use of different forms of estimating, clear descriptions of the calculations used to produce indices, and clinical acceptability should be emphasized when assessing reliability and method comparison studies.
(Ewha Med J 2017;40(1):9-16)

Received December 29, 2016
Accepted January 4, 2017

Corresponding author
Kyoung Ae Kong
Clinical Trial Center, Ewha Womans University Medical Center, 1071 Anyangcheon-ro, Yangcheon-gu, Seoul 07985, Korea
Tel: 82-2-2650-2069, Fax: 82-2-2650-6141
E-mail: kkong@ewha.ac.kr

Key Words
Validation studies; Reliability; Reproducibility of results; Agreement; Method comparison

서 론

임상 또는 의학 연구에서 위험요인을 선별하거나 질병을 진단하거나 또는 환자의 예후를 추정하기 위해 측정을 하게 되며, 적절한 측정 방법, 검사법의 선택은 임상적 결정을 내리는 과정과 연구의 질을 보장하는 데 있어 필수적이다. 검사법은 주로 타당도와 신뢰도의 측면에서 평가되곤 한다. 타당도(validity)는 검사법이 측정하고자 하는 바를 제대로 반영하는 능력이나 실제 모수의 값을 정확하게 관찰하는 능력을 의미하며 정확도(accuracy)와 동일한 의미로 사용된다[1-3]. 신뢰도(reliability)는 검사 시기, 실험실, 평가자 등 측정 조건과 상관없이 검사결과가 얼마나 일관되게 나타나는지의 일관성(consistency), 또는 측정 오차(measurement error)가 없는 것을 말하는 것으로, 반복성(repeatability), 재현성(reproducibility), 일치도(agreement, concordance) 등이 동일한

의미로 사용되곤 한다[1,3]. 반복성은 동일 대상자를 동일 조건(동일 도구나 방법, 동일한 평가자, 측정하고자 하는 대상에 변화가 없을 짧은 시간 차이)에서 반복 측정한 값들에서의 변동(variation)를 의미하며, 따라서 반복성 연구에서 동일 대상자의 측정값 변이는 측정 과정 그 자체에 기인한 오차에만 기인한다[4]. 재현성은 다른 조건 하에서 동일 대상자에 대한 측정값의 변이를 말하는데, 다른 조건이란 다른 검사법이나 도구, 다른 평가자, 측정하고자 하는 변수에 무시할 수 없는 변화가 일어날 수 있을 정도의 시간 경과 등을 의미한다. 반복성과 재현성은 평가 방법이 동일하며, 다만 재현성을 평가하려면 먼저 반복성이 평가되고 인정되어야 한다[5]. 또한 반복성과 평가자간 재현성과 같은 신뢰도는 검사법 비교나 타당도 평가의 전제조건이다[1,5]. 검사법을 평가하는 또 다른 접근은 검사법 비교(method comparison)이다[3]. 검사법이 새롭게 개발되면 먼저 기존에 사용되고 있는 검사법과 비교를 통해 평가 받게 되는데, 표준 검사법은 황금기준 검사법이라고 불리기도 하지만 오류 없이 측정해 낸다는 뜻은 아니다[6]. 기존 검사법과 새로운 검사법 중 어느 것이 더 올바르게 측정해 내는지 분명하지 않을 수 있으며, 대개는 두 검사법 간 일치의 정도를 평가하게 된다. 따라서 검사법 비교는 재현성 연구의 일종이라고 할 수 있다. 검사법 비교를 타당도 연구의 특별한 종류로 언급한 경우도 있기는 하지만[3], 민감도, 특이도, ROC (receiver operating characteristics) 곡선 등을 이용하는 진단검사법의 타당도 연구와는 사용하는 통계 방법이나 해석이 다르고 오히려 신뢰도 연구에서와 유사한 방법들을 사용하는 경우가 많다. 여기서는 검사법 비교와 신뢰도 평가에 흔히 사용되는 통계적 방법과 지표들을 살펴보고자 한다.

본 론

1. 검사법의 측정값이 연속형 변수인 경우

1) 급내상관계수

급내상관계수(intraclass correlation coefficient, ICC) 또는 신뢰도 계수(reliability coefficient)는 반복성과 재현성을 평가하는데 매우 흔하게 사용되는 지표로, 측정값들의 총 변동 중 개인간 변동에 의해 야기된 부분에 대한 추정치이다[2]. 이를 구하는 공식은 아래와 같다.

$$\mathrm{ICC} = \frac{V_b}{V_T} = \frac{V_b}{V_b + V_e}$$

V_T: 총 변동(total variance), V_b와 V_e의 총합

V_b: 일반적인 신뢰도 연구에서는 개인간 변동(variance between individuals)

V_e: 개인내 변동(variance within individuals), 원하지 않은 변동, 오차, 동일 대상에 대한 여러 측정값들 간 분산의 추정치

ICC는 0 (전혀 일치하지 않음)부터 1 (완벽하게 일치함) 사이의 값을 갖는다. Shrout와 Fleiss [7]는 분산분석의 종류(일원배치 혹은 이원배치), 평가자 효과(평균 측정치의 차이) 고려 여부, 분석 단위(개별 또는 평균 측정치)에 따라 어떤 ICC를 선택해야 할지 제시하였다. 첫 두 가지는 통계적 모형 선택과 관련되며, 두번째와 세번째는 연구 결과의 용도와 관련된 것이다.

전형적인 평가자간 신뢰도 연구에서는 n명의 대상자가 k명의 평가자에 의해 독립적으로 평가를 받게 되는데, 다음과 같은 세가지 경우로 분류할 수 있다: (1) 각 대상자가 평가자 집단으로부터 무작위 추출된, k명으로 구성된 서로 다른 평가단(평가자 세트)에 의해 평가; (2) 평가자 집단으로부터 무작위 추출된 k명의 평가자 각각이 n명의 대상자 모두를 평가; (3) 연구의 관심 대상인 k명의 평가자가 있으며, 이들 각각이 n명의 대상자 각각을 평가. 이 중 (1)의 경우는 일원배치 분산분석으로 분석한다—Shrout와 Fleiss [7]의 ICC(1,1). (2)의 경우는 해당 결과를 인구집단내의 다른 평가자들에게까지 일반화하는 목적이 있으며, 평가자를 임의효과로 취급하는 이원배치 변량(임의)효과 모형(two-way random effects model)으로 분석한다—ICC(2,1). 측정값의 변동에서 평가자의 효과를 고려하며, 이 때의 ICC는 신뢰도 연구에서 일반적으로 목표하는 일치도(agreement)로서, 평가자들의 교환 가능성이라고도 할 수 있다. (3)의 경우는 연구의 관심이 단 한 명의 평가자 또는 고정된 k명의 평가자로서, 평가자를 고정효과로 가정하는 이원배치 혼합효과 모형(two-way mixed effects model)을 이용한다—ICC(3,1). 이 모형에서는 평가자에 의한 변동을 고려하지 않으며, 산출된 ICC는 평가자간 일관성(consistency)으로 해석한다. 대부분의 신뢰성 연구에서 연구자들은 일관성이 아니라 일치도에 관심이 있고 그들의 평가 척도가 여러 평가자에게 사용되기를 바랄 것이므로 일반적으로는 이원배치 변량효과 모형이 적절하다—ICC(2,1).

신뢰도의 분석 단위는 실제로 측정값들을 산출하는 정황과 관련되는데, 개별 측정값들이 아니라 여러 평가자가 측정한 값들의 평균을 분석단위로 이용하는 경우에는 평균 ICC를 사용하게 된다—ICC(1,n), ICC(2,n), ICC(3,n). 일반적으로 개별 ICC보다 평균 ICC가 더 높지만, 이 검사를 이용하는 실제 상황에서 여러 평가자의 평균값을 이용하는 것이 아니라면 평균 ICC를 사용하는 것은 적절하지 않다.

일치도의 지표로서 ICC는 상관계수보다 더 좋은 지표로 여겨지는데, 상관관계와 측정값 간 바이어스의 정보를 둘 다 포함하고 있기 때문이다[2]. 두 번의 측정값들 간에 구조적인 차이가 있는 경우(예를 들어 두 번째 측정값이 첫 번째 측정값보다 항상 0.5만큼 낮게 측정된다거나 첫 번째 측정값의 80% 크기로 측정되는 경

우)에도 두 측정값들이 선형적 관련성을 나타내는 직선에 가깝게 모인다면 상관계수는 매우 높게 나타나게 되지만, ICC는 이와 같은 바이어스를 반영하여 상관계수보다 낮다(Fig. 1).

ICC는 연구집단에서의 측정값들의 범위에 영향을 받기 때문에 해석할 때 주의가 필요하다. 예를 들어 연구집단의 측정값들이 전체적으로 다 높은 수준이고 범위가 작다면 개인내 변동(오차)에 비해 상대적으로 개인간 변동이 작고 ICC가 낮다. 이와 같이 표본의 특성이 반영되므로 서로 다른 연구집단의 ICC는 비교하기 어렵다[5]. 또한 0.40미만은 좋지 않음(poor), 0.4–0.6은 보통(fair), 0.6–0.75는 좋음(good), 0.75–1.00은 매우 좋음(excellent) 등으로 분류되기도 한다[8,9]. 하지만 절대적인 기준은 없고, 단위가 없는 지표이므로 오차의 실제 크기와 상관이 없어서 연구에서 나타난 오차가 임상적으로 받아들일 수 있는 수준인가 하는 관점에서 해석하기 어렵다.

ICC는 반복성과 평가자내 또는 평가자간 재현성 연구에 많이 이용되며, 검사법 비교 연구에서는 검사법간 직접 비교에 사용되는 경우도 있긴 하지만[10,11] 주로 두 검사법의 신뢰도를 각각 ICC로 제시하고 대조해 보는 용도로 사용되곤 한다. 때로는 동일한 대상으로부터 얻어진 두 검사법의 평가자내, 평가자간 ICC를 직접 비교하기도 하는데, 이때는 Fisher의 Z-검정, Konishi-Gupta의 수정 Z-검정 등을 할 수 있으나[12], 흔히 사용하는 통계 패키지 내에 들어있는 기능은 아니다.

2) Bland-Altman 그림과 일치한계값들

Bland-Altman 그림[6,13,14]은 동일 대상에 대한 두 세트의 측정값에서 각 측정값의 짝마다 평균과 차이(mean of and difference between each pair of measurements)를 계산한 다음 평균을 x축, 차이를 y축으로 하는 산점도인데, 반복성과 재현성 평가에서 뿐만 아니라 서로 다른 두 검사법에 의한 측정값들 간의 불일치(disagreement) 양상을 살펴보기에 매우 유용해서 검사법 비교 연구에서 많이 사용되고[11,15] 권고되는 방법이다[3,4,6,13,14,16,17]. 두 측정값 중 하나가 황금-기준 검사법에 의한 참값인 경우에는 그 값을 x축으로 하기도 한다. Bland-Altman 그림에는 일반적으로 x축과 평행한 세 개의 가로선을 표시하는데, 불일치의 정도를 요약하는 값들이라고 할 수 있다. 가운데 가로선은 평균 차이(mean difference, $\bar{d}$), 즉 측정값의 짝 간 차이의 평균(mean of the differences between measurements)을 나타내는데, 0으로부터 이 값까지의 거리는 두 검사법(또는 평가자) 간 바이어스의 추정치라고 할 수 있다. 변동은 측정값 간 차이의 표준편차(standard deviation of the differences, s_d)로부터 추정하게 된다. 먼저 차이의 표준편차의 1.96배 값($1.96s_d$)을 구하는데, 동일한 방법으로 동일한 대상에서 얻어진 측정값들인 경우에는 이 값을 반복성 계수(repeatability coefficient)라고 한다[4-6]. 이 값을 평균 차이에 더하고 뺀 값($\bar{d}\pm1.96s_d$)을 구하여 95% 일치한계값들(limits of agreement, LOA)이라고 한다.

95% 일치한계 상한값(upper LOA): $\bar{d}+1.96s_d$
95% 일치한계 하한값(lower LOA): $\bar{d}-1.96s_d$

평균 차이를 나타내는 가로선의 위와 아래에 있는 가로선은 95% LOA 상한값과 하한값을 나타내며, 측정값 간 차이들이 정규분포를 따른다면 차이의 대략 95%는 LOA 상한과 하한 사이에 존재하게 된다. 주의할 점은 평균 차이와 95% LOA는 측정값의 전 범위에 걸쳐 바이어스와 변동이 균일한 경우에만 의미가 있다는 것이다(뒷부분의 Fig. 2와 Fig. 3 해석에서 추가 설명). 평균 차이와 95% LOA는 표본으로부터의 추정치이므로 표준오차나 95% 신뢰구간과 같은 정밀성의 정보를 같이 제공하게 되는데, 계산 방법은 다음과 같다[6,13,16].

평균 차이의 표준오차: 측정값 간 차이의 표준편차를 표본수(n)의 제곱근으로 나눈 값($s_d/\sqrt{n}$)

LOA의 표준오차: $\sqrt{\left(\frac{1}{n}+\frac{1.96^2}{2(n-1)}\right)s_d^2}$, 평균 차이의 표준오차의 약 1.71배($\sqrt{2.92}\times s_d/\sqrt{n}$).

95% 신뢰구간 상하한값: 해당 추정치(평균 차이,LOA 상하한값)$\pm t_{n-1,.975}$)×해당 표준오차

때로는 각 검사법(또는 각 평가자)에 대해 반복적으로 측정을 하고, 반복 측정값의 평균을 각 검사법에 대한 측정치로 사용하는

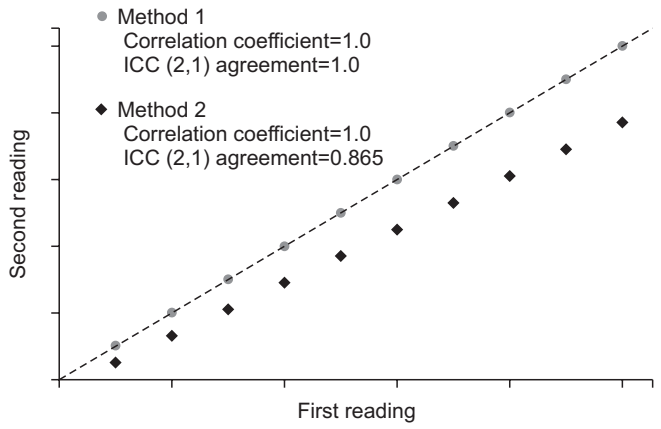

Fig. 1. Intraclass correlation coefficient and Pearson's correlation coefficient as indices for intra- or interobserver reliability. ICC, intraclass correlation coefficient; correlation coefficient, Pearson's correlation coefficient.

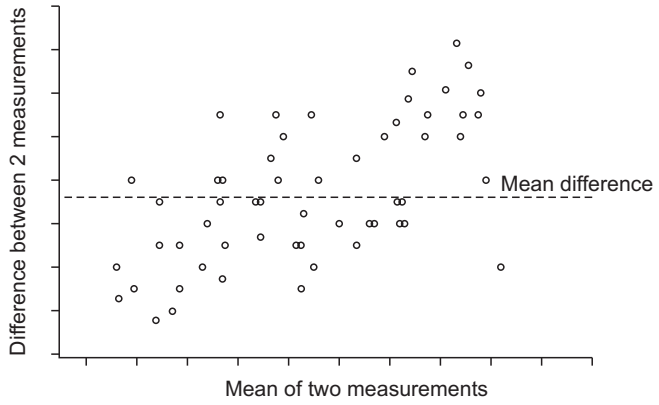

Fig. 2. Graphical presentation of agreement. A case where the greater magnitude of measurements has the greater difference.

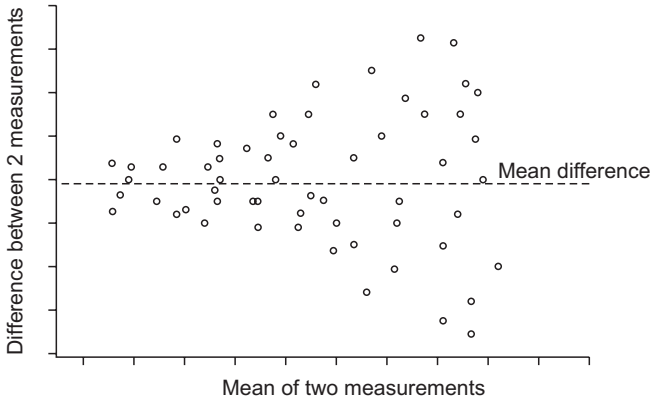

Fig. 3. Graphical presentation of agreement. A case where an increase in the variability of the differences is based on an increase in the magnitude of measurements.

경우가 있다. 그러나 이 값들을 이용해서 검사법을 비교하면 검사법간 차이의 표준편차가 과소 추정되므로(반복 측정 오차의 효과 중 일부가 제거되었기 때문), 아래와 같이 수정된 표준편차를 구한 후 이를 LOA 계산에 사용해야 한다.

$$\hat{\sigma}_d = \sqrt{s_{\bar{d}}^2 + \left(1 - \frac{1}{m_x}\right) s_{xw}^2 + \left(1 - \frac{1}{m_y}\right) s_{yw}^2}$$

$\hat{\sigma}_d$: 방법 x와 방법 y 간 차이의 수정된 표준편차

$s_{\bar{d}}^2$: 방법 x의 반복측정 평균과 방법 y의 반복측정 평균의 차이의 분산

m_x, m_y: 방법 x와 방법 y의 각 대상자 당 관찰 개수(반복측정 횟수)

s_{xw}^2, s_{xw}^2: 방법 x와 방법 y 각각에서의 개체내 변이 - 각 방법별로 측정값들을 종속변수로 하고 각 대상자를 요인으로 하는 일원분산분석을 했을 때 평균제곱오차(mean square error)로 추정

이렇게 각 방법 별로 반복 측정값들의 평균을 이용해서 얻어진 방법간 차이의 평균에 수정된 표준편차($\hat{\sigma}_d$)의 1.96배 값을 더하고 빼서 95% LOA를 구한다. 이 LOA의 표준오차와 95% 신뢰구간을 구하는 방법은 Bland와 Altman [6,13,16]의 논문에 소개되어 있다.

Bland-Altman 그림과 평균차이, LOA는 대략 세 단계로 살펴보게 된다. 먼저 x축 값(참값이나 측정값 짝의 평균)의 크기에 따른 불일치의 분포 양상을 살펴보게 된다. Fig. 2에서는 값이 클수록 짝진 두 측정값 간의 차이가 양의 방향으로 커지고, Fig. 3에서는 값이 클수록 짝진 두 측정값 간 차이의 변동이 커진다. 이와 같이 불일치 정도가 측정값의 크기와 관련이 되는 경우에는 평균 차이와 LOA를 제시하는 것은 의미가 없으며[4,5], 그림과 함께 측정값과 관련된 불일치의 양상을 기술하는 것이 더 적절할 수 있다. 또한 Fig. 3에서와 같이 값이 커질수록 오차가 커지는 경우에는 측정값들을 로그 변환하여 분석하는 방법을 고려해 볼 수 있고 [4,6,13,14], Fig. 2와 같은 양상을 나타내는 경우에는 차이(불일치)를 측정값의 크기에 따른 함수로 모형화 하는 회귀분석적인 접근 방법을 고려해 볼 수 있다[6].

측정값 짝 간 차이의 변동이 측정값들의 전 범위에서 일정하다면, 다음 단계로는 두 검사법(또는 평가자) 간 일치/불일치를 나타내는 지표들을 정량화하여 제시하고 해석한다. 평균 차이는 두 검사법(또는 평가자) 간 바이어스(구조적 차이)가 있는지, 한 방법이 다른 방법에 비해 평균적으로 과다 혹은 과소 추정하는 경향이 있는지 알려준다[4]. 반복성 연구에서는 평균 차이가 0일 것을 가정하고 있으며 그렇지 않은 경우 사실상 반복 측정치가 아닐 가능성을 검토해 보게 되는 반면, 검사법 비교나 재현성 연구에서는 바이어스가 있을 수 있다고 인정하고 이를 평균 차이로 요약하는 것이다. 평균 차이가 0에 매우 가깝다면 바이어스의 가능성은 적고, 한 방법이 다른 방법보다 과다 혹은 과소 추정하는 경향은 없다고 할 수 있다. Fig. 4는 서로 다른 두 영상의학적 검사법 A와 B로 얻어진 폐결절의 크기 평균과 차이에 대한 Bland-Altman 그림이다. 방법 A로 측정된 결절의 크기는 방법 B로 측정된 결절의 크기보다 평균적으로 0.25 mm 작은데, 아마도 이 정도의 과소추정은 임상적으로 받아들일 수 있다고 해석하게 될 것이다. 때로는 바이어스를 평가하기 위해 평균 차이가 0이라는 귀무가설을 가지고 짝진 t-검정이나 일표본 t-검정을 시행하고, P값이 큰 경우에는 바이어스의 근거가 없다고 언급하곤 한다. 그러나 이러한 검정들은 평균 차이가 0에 매우 가까운 경우뿐만 아니라 방법 간의 무작위 오차(차이들의 변동)가 큰 경우에도 귀무가설을 기각하지 못하며 [3], 두 방법간 바이어스가 크지 않다는 오해를 일으키기 쉽다. 또한 평균 차이는 두 방법간 차이가 평균적으로 0에 가까운지, 즉

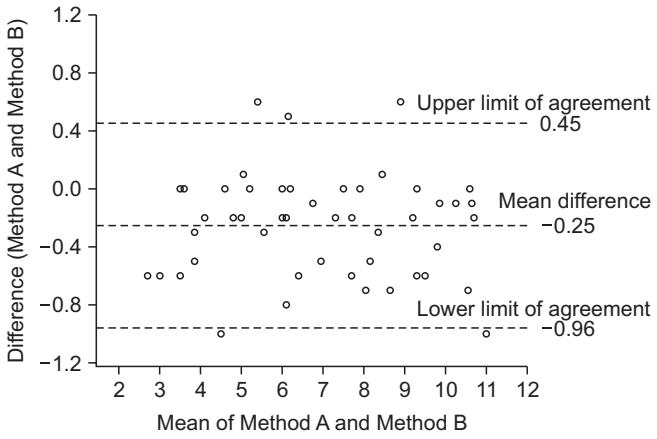

Fig. 4. Measurements of pulmonary nodule size using two radiological methods (shown is a Bland-Altman plot).

두 방법 중 하나가 평균적으로 과소/과다 추정하는 경향 만을 검토하는 것이므로 평균 차이에 대한 검정이 유의하지 않다고 해서 두 방법의 측정값이 대부분 일치한다거나 두 방법을 교환해서 사용할 수 있다는 의미가 아니라는 점을 유의해야 한다. 이런 면에서 신뢰성 연구나 두 검사법을 비교하는 연구에서 짝진 t-검정을 시행하고 해석하는 데에는 특별한 주의가 필요하며, 이 결과만을 단독으로 제시하지 말고 다른 평가 방법들의 결과와 함께 제시해야 한다[18].

마지막으로는 95% LOA의 값과 범위를 평가하게 된다. 검사법에 따른 측정값 간 차이들이 정규분포를 따른다면 차이의 대략 95%는 95% LOA 상한과 하한 사이에 있을 것이므로, LOA 값들과 범위가 임상적으로 받아들여질 만 하다면 두 검사법은 교환 가능하다고 할 수 있다. Fig. 4의 LOA 상한, 하한값은 0.45와 -0.96 mm이고 두 방법간 차이의 95%는 이 1.5 mm (=0.45–(-0.96)) 범위 이내에 존재할 것이므로, 이 값들과 범위가 임상적으로 수용 가능하다면 A 방법과 B 방법은 서로 바꾸어 사용할 수 있다. 때로는 임상적으로 의미있는 크기나 기준이 잘 알려져 있지 않은 생체지표를 서로 다른 방법으로 측정하는 경우가 있는데, 이런 경우에는 불일치의 크기를 측정값들의 범위와 비교해서 해석하기도 한다. Fig. 4에서 측정값들의 임상적 의미가 명확하지 않다면 LOA의 폭인 1.5를 측정값의 범위인 8.5 (대략 2.5–11)와 비교해서 두 검사법간 차이의 변동(오차의 범위)이 받아들일만한 수준인지를 판단해 볼 수 있다.

Fig. 3에서와 같이 측정값이 클수록 변동이 커지는 경우에는 자료를 로그 변환하면 측정값과 변동이 더 이상 관련이 없게 되는 경우가 많다. 이런 경우에는 로그 변환 자료로부터 구한 LOA를 역변환하여 검사법 간 비의 한계(limits for the ratio of the actual measurements)를 제시할 수 있다[6,13]. 먼저 모든 측정값들을 로그 변환한 다음, 각 짝마다 평균과 차이를 구하여 이를 x축과 y축으로 하는 산점도를 그리고 측정값의 크기에 따른 변동의 크기가 균일한 것을 확인한다. 로그 변환된 값들의 차이의 평균과 표준편차로부터 LOA를 구한 뒤 이를 역변환하여 비의 한계를 구한다. 만약 로그 변환된 자료로부터 차이의 평균과 LOA 상한, 하한값이 0.05, -0.07, 0.17 이라면 역대수값(antilog)은 1.05, 0.93, 1.19 이다. 로그 변환된 두 값 차이의 역대수값은 비이므로, '대략 95%의 경우에 검사법 A를 이용한 측정치는 검사법 B를 이용한 측정치의 0.93–1.19배 사이에 있을 것이다' 또는 '대부분의 경우 검사법 A에 의한 측정값은 검사법 B보다 7% 작거나 19% 큰 정도 이내의 차이가 있을 것이다' 라고 해석할 수 있다. 다른 예로, 로그 변환값 차이의 평균과 LOA의 역대수값이 1.16, 1.11, 1.22 이라면, 대부분의 경우 검사법 A에 의한 측정값은 검사법 B보다 11%–22% 클 것이라고 해석할 수 있으며 1.16을 변환계수(conversion factor)로 이용해서 검사법 A에 의한 측정값들을 먼저 1.16으로 나누면 두 검사법간 일치도는 훨씬 높아질 것이다. 로그 변환 값들의 차이를 이용하지 않고 각 짝에서 직접 두 측정값의 비를 계산해서 이 값들의 평균과 표준편차로부터 LOA를 구할 수도 있으며, 그림을 그릴 때 y축을 평균에 대한 백분율로 나타내는 경우도 있다[6].

3) 변동계수

변동계수(coefficient of variation, CV)는 실험실적인 연구나 생화학적 분석에서 신뢰도 또는 측정 오차의 지표로 사용된다. 검사법 비교에는 거의 사용되지 않는다. 일반적으로 CV는 자료의 표준편차를 평균으로 나누고 100을 곱하는 것이라고 말하지만, 신뢰도 연구에서 CV를 계산하는 데에는 여러 방법이 있다. 가장 단순한 방법은 각 개인별로 측정값들의 CV (individual CV)를 내고, 개인별 CV로부터 평균 CV (mean CV)를 계산하는 방법이지만, CV를 사용하는 데에는 여러 가지 고려할 점과 제약이 있다[3]. 자료가 음의 값을 가질 수 있거나 측정 척도의 중간에 0 값이 있는 경우에는 CV가 의미 없을 수 있고 이를 사용하는 것은 부적절하다. 실제적인 의미를 이해하는 데에도 주의가 필요하다. 예를 들어, CV가 10%라는 것은 측정간 변이(차이)를 나타내는 모든 값이 항상 평균의 10% 이내에 있다는 것을 의미하는 것이 아니라 자료의 정규분포를 가정했을 때 차이의 68%가 자료 평균의 10% 이내에 있다는 의미이고, 나머지 32%에 대해서는 언급하지 않은 것이다. 특히 개인별 CV로부터 계산된 평균 CV는 측정간 변동을 실제보다 과소평가할 수 있고 전체가 아니라 평균적인, 즉, 표본에 있는 사람들의 50%에서의 변동만을 반영할 수 있어 분석의 목표가 되는 경우는 드물다. CV는 불일치의 정도가 측정값의 크기에 따라 커진다는 것을 가정하고 있고 그러한 자료에 적용된다. 따라서 로그 변환된 자료에서 각 대상자를 변량효과로 처리하는 분산

분석(random-effects model one-way ANOVA)을 하고, 개체내 평균제곱(평균제곱오차 항 mean square error term)에 기반해서 원 척도에서의 CV를 계산하는 것이 더 적절한 방법으로 제시된다 ($CV=\sqrt{\text{평균제곱오차}} \times 100\%$) [3,9]. 일반적으로 CV가 20% 미만인 것을 바람직한 것으로, 30% 이상은 적절하지 않은 것으로 판단한다[9].

2. 진단방법의 측정값이 범주형 변수인 경우

1) 일치율

측정 결과가 범주형 변수이고 관찰값의 세트가 둘인 경우(두 검사법이나 두 명의 평가자에 의한 측정, 시간을 두고 두 번 반복 측정한 경우 등), 관찰값들의 짝 중 판정이 일치하는 짝의 비율을 일치율(percent agreement)이라고 하며 Table 1에서 다음과 같이 계산된다[1,2].

percent agreement (%)=100×(a+d)/(a+b+c+d)

일치율은 매우 간단하고 판정 범주의 개수가 세 개 이상인 경우에도 쉽게 산출할 수 있다는 장점이 있지만, 연구 집단의 양성율(질병 유무를 측정해 내는 검사인 경우는 유병률)이 낮은 경우에는 두 검사법(평가자) 모두 음성으로 판정하는 음성-음성 결과가 차지하는 비율이 높아져서 일치율이 과다하게 높게 추정되는 단점이 있다[2].

2) 양성 일치율

인구집단에서 측정하고자 하는 상태의 유병률이 매우 낮거나 높은 경우에 신뢰도 지표로서의 일치율의 단점을 극복한 두 가지 양성율이 있다[2]. 첫째는 양성 일치율(percent positive agreement)로, 두 평가자 모두 양성으로 판정한 관찰값의 수를 두 평가자가 각각 양성으로 판정한 관찰값 수의 평균으로 나누어 계산한다.

$$\text{Percent positive agreement} = 100 \times a/\left(\frac{(a+c)+(a+b)}{2}\right) = 100 \times 2a/(2a+b+c)$$

Table 1. Agreement between observers A and B on binary measurements

Observer A	Observer B		Total
	Positive	Negative	
Positive	a	b	a+b
Negative	c	d	c+d
Total	a+c	b+d	N

다른 하나는 Chamberlain 양성일치율(Chamberlain's percent positive agreement)로, 두 평가자 모두 양성으로 판정한 관찰값의 수를 적어도 한 평가자가 양성으로 읽은 관찰값의 수로 나누어 계산한다.

Chamberlain's percent positive agreement=100×$a/(a+b+c)$

3) 카파 통계량과 가중 카파 통계량

일치율은 두 평가자의 판정이 우연히 일치하는 부분을 고려하지 못한다는 단점이 있다. 평가자가 검사대상을 판정할 때 무작위로 절반은 양성으로, 절반은 음성으로 판정하는 경우를 생각해 보면, 두 평가자가 서로 독립적으로 판정을 했다고 할지라도 우연에 의해 대략 절반 정도는 두 평가자의 판정이 일치하게 될 것이다. 카파 통계량(Cohen's kappa statistic)은 이와 같은 우연에 의한 일치를 감안한 일치도로, 다음과 같이 계산된다[2].

$$\text{Cohen의 kappa} = \frac{\text{우연에 의하지 않은 관찰된 일치율}}{\text{우연에 의하지 않은 최대 일치율}} = \frac{\text{관찰된 일치율} - \text{우연에 의한 일치율}}{\text{최대 일치율}(1.0) - \text{우연에 의한 일치율}}$$

위의 식에서 우연에 의한 일치율은 각 평가자가 무작위로 판정을 했다고 가정했을 때 기대되는 일치율로, 각 평가자가 양성으로 판정한 비율에 근거하여 구하게 된다. 예를 들면 Table 1에서 두 평가자 모두 양성으로 판정한 칸(a)에서 기대되는 우연에 의한 일치율(즉, 평가자 둘 다 우연에 의해 양성으로 판정할 확률)은 평가자 A가 양성으로 판정한 비율과 평가자 B가 양성으로 판정한 비율의 곱(두 판정이 독립이라는 가정 하에서의 결합확률)으로 구하고, 전체 우연에 의한 일치율은 판정이 일치하는 대각선 상에 있는 모든 칸의 우연에 의한 일치율을 합해서 얻는다.

$$\text{우연에 의한 일치율} = \left[\frac{(a+b)}{n} \times \frac{(a+c)}{n}\right] + \left[\frac{(c+d)}{n} \times \frac{(b+d)}{n}\right] = [(a+b)(a+c)+(c+d)(b+d)]/n^2$$

카파 통계량은 판정이 두 범주가 아니라 더 많은 범주로 이루어진 경우에도 일치를 나타내는 모든 칸들로부터 관찰 일치율과 우연에 의한 일치율을 구해서 계산하면 된다. 다만 판정의 범주가 많아지면 카파값은 낮아진다[5].

카파 통계량은 계산상으로는 -1부터 1까지 가능하지만 0 미만의 값은 우연보다도 낮은 일치를 나타내므로 실제적인 범위는 0 (전혀 일치하지 않음)에서 1 (완벽한 일치)이다. Landis와

Koch [19]는 0.80 보다 크면 거의 완벽한 일치(almost perfect), 0.61–0.80은 상당한 크기의 일치(substantial), 0.41–0.60은 적당한 크기의 일치(moderate), 0.21–0.40은 어느 정도의 일치(fair), 0.0–0.20은 약간의 일치(slight), 0 이하는 일치도 나쁨(poor)으로, Fleiss [20]는 0.75 이상은 매우 좋음(excellent), 0.4–0.75는 어느 정도 일치에서 좋음 사이(fair to good), 0.40 미만은 좋지 않음(poor)으로, Altman [21]은 0.8 이상은 매우 좋음(very good), 0.6–0.8은 좋음(good), 0.4–0.6는 적당(moderate), 0.2–0.4는 어느 정도(fair), 0.2 미만은 좋지 않음(poor)으로 분류하였다[2,5]. 카파 통계량에는 대부분 P값을 붙이지 않는 데 카파값이 0 이라는 귀무가설(동일 대상자에 대한 측정인데 전혀 일치하지 않음)에 대한 검정은 의미가 없어서 일반적으로 시행하지 않기 때문이며[5], 이는 ICC에 대해서도 마찬가지이다.

판정이 여러 범주로 이루어진 경우, 어느 정도 가까운 범주로 판정해 낸 경우는 부분적으로 일치한 것으로 인정해줄 수 있기도 하고 어떤 범주 간의 불일치는 특별히 심각한 것일 수 있는데, 이와 같은 상황에 대한 고려하여 불일치의 정도에 따라 가중치를 부여해서 계산하는 것이 가중 카파 통계량(weighted kappa)이다[2]. Table 2에 가중치를 부여하는 예를 제시하였다. 두 판정이 일치하는 대각선 상의 칸들(a, f, k, p)은 가중치를 1로 부여하고, 불일치하되 인접하는 판정인 경우(b, g, l, e, j, o)는 가중치를 0.75로, 한 칸 더 떨어진 거리로 판정하는 경우(c, h, i, n)는 가중치를 0.5로 부여한다면, 관찰된 일치율과 우연에 의한 일치율은 각 칸의 관찰 빈도 또는 우연에 의한 일치 빈도에 가중치를 곱하는 다음과 같은 식으로 계산한다.

$$\text{관찰된 일치율} = \begin{bmatrix} (a+f+k+p)\times 1.0 \\ +(b+g+l+e+j+o)\times 0.75 \\ +(c+h+i+n)\times 0.5 \end{bmatrix} \Big/ n^2$$

Table 2. Agreement between methods A and B on measurements with four-category results

Method A	Method B				Total
	Definite	Probable	Possible	Absent	
Definite	a (1.0)	b (0.75)	c (0.5)	d (0.0)	A1
Probable	e (0.75)	f (1.0)	g (0.75)	h (0.5 or 0)	A2
Possible	i (0.5)	j (0.75)	k (1.0)	l (0.75 or 0)	A3
Absent	m (0.0)	n (0.5 or 0)	o (0.75 or 0)	p (1.0)	A4
Total	B1	B2	B3	B4	N

Number in parentheses indicates the weight used for calculation of the weighted kappa.

우연에 의한 일치율

$$= \begin{bmatrix} \{(A1\times B1)+(A2\times B2)+(A3\times B3)+(A4\times B4)\}\times 1 \\ +\begin{Bmatrix}(A1\times B2)+(A2\times B3)+(A3\times B4)\\ +(A2\times B1)+(A3\times B2)+(A4\times B3)\end{Bmatrix} \\ \times 0.75 \\ +\{(A1\times B3)+(A2\times B4)+(A3\times B1)+(A4\times B2)\}\times 0.5 \end{bmatrix} \Big/ n^2$$

$$\text{Weighted kappa} = \frac{\text{관찰된 일치율} - \text{우연에 의한 일치율}}{1.0 - \text{우연에 의한 일치율}}$$

질병이 아니라고 잘못 판단하는 것이 심각한 문제라고 생각한다면, 똑같은 간격을 두고 떨어져 있다고 할지라도 한 검사법에서 질병이 아니라고 판정한 경우(d, h, l, m, n, o칸)는 가중치를 0을 줄 수도 있다. 가중치는 자료가 사용될 실제 상황을 고려했을 때 이 불일치가 얼마나 심각한 문제인가에 대한 연구자의 인식에 기초해서 부여하게 되는데 이러한 인위성은 가중 카파의 약점 중 하나이며, 특히 연속형 변수를 여러 개의 범주로 만든 순위형 변수의 경우에 문제가 될 수 있다[2].

카파 통계량을 해석할 때는 몇 가지 주의해야 할 점이 있다. 첫째로 각 집단의 실제 양성 유병률(prevalence of true positivity)이 0이나 1에 가까우면 카파값은 작아지고 0에 가까워 지는 경향이 있기 때문에 서로 다른 집단의 검사법 신뢰도를 비교할 때는 주의를 요한다. 또한 두 평가자(또는 두 검사법)가 판정한 양성률이 비슷할 때보다 서로 다를 때의 카파값이 더 큰 경향이 있다. 따라서 카파 통계량은 일치율과 같은 일치도의 다른 측도들을 함께 제시하고, 해석을 할 때 해당 조건의 유병률 및 관찰자들간 양성 유병률이 얼마나 비슷한지를 고려할 필요가 있다.

세 명 이상의 관찰자가 평가를 한 경우 또는 세 번 이상의 반복 측정이 이루어진 경우(multiple ratings)에는 Fleiss [22]의 카파 통계량을 이용할 수 있다. Fleiss의 카파는 Cohen의 카파와는 달리 모든 대상자가 동일한 평가자에 의해 평가를 받아야 한다는 가정을 갖고 있지 않으며, 각 대상자가 서로 다른 관찰자에 의해 평가받는 경우를 가정한다[23]. 두 명의 관찰자가 평가한 경우에 Cohen의 카파와는 다르다는 지적을 받기도 하지만, 일반적으로는 세 명 이상의 평가자가 평가한 경우에 대한 카파의 확장으로 Fleiss의 카파를 사용하고 있으며, STATA 등 많이 사용하는 통계 패키지로 분석할 수 있다.

결 론

지금까지 의학 문헌에서 검사법 비교와 신뢰도 평가에 흔하게 사용되는 통계 방법과 지표들을 살펴보았다. 신뢰도나 검사법 비교 연구는 가설 검정을 사용하는 일이 적고 결과 해석에 주관적이고 기술적인 면이 많아 연구자들에게 어렵게 생각될 수 있다. 각 신뢰도 평가 방법의 장점과 제한점을 고려하여 적절한 몇 가지를

같이 사용하고 제시하는 것이 권고된다[18]. 지표들의 산출 방법은 자료 변환, 반복 측정한 값들 중 어느 값을 사용했는지, 사용한 통계 분석 절차 같은 면까지 좀 더 정확하게 기술하고, 검사법의 활용과 관련된 임상적 의의를 고려하여 해석하는 것이 필요할 것으로 생각된다. 신뢰도 연구에서 주의할 점을 다시 몇 가지 언급한다면, 반복성은 재현성의 전제조건이며 반복성과 평가자간 재현성은 검사법간 비교의 전제 조건일 수 있으므로, 전제 조건에 해당하는 신뢰도가 먼저 확인되어야 한다. 반복성 평가에서 얻어진 여러 반복 측정값들을 평균을 내서 사용하면 검사법간 비교나 재현성 평가에서 변동을 잘못 추정할 수 있으므로 평균을 사용하지 않거나 평균을 사용할 때 적절한 통계 방법을 사용해야 한다. 서로 다른 연구에서 얻어진 신뢰도는 직접 비교하기 어렵다. 마지막으로, 각 관찰은 서로 독립으로 이루어져야 한다. 매우 당연하고 쉬운 일인 것 같지만 기존 임상자료를 재구성하여 신뢰도 연구를 하거나 검사법 간 비교를 하는 경우에는 평가자가 해당 영상이나 환자를 기억하거나 관련 정보를 갖고 있을 수 있으므로 각 관찰에서 독립성을 확보하기 위하여 주의를 기울여야 할 것이다.

References

1. Korean Society for Preventive Medicine. Preventive medicine and public health. 2nd ed. Seoul: Gyechuk Munwhasa; 2013.
2. Szklo M, Nieto FJ. Epidemiology: beyond the basics. 2nd ed. Sudbury, MA: Jones and Bartlett Publishers; 2007.
3. Atkinson G, Nevill AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. *Sports Med* 1998;26:217-238.
4. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables. *Ultrasound Obstet Gynecol* 2008;31:466-475.
5. Petrie A, Sabin C. Medical statistics at a glance. 3rd ed. Chichester, UK: John Wiley & Sons; 2009.
6. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;8:135-160.
7. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 1979;86:420-428.
8. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess* 1994;6:284-290.
9. Rosner B. Fundamentals of biostatistics. 7th ed. Boston, MA: Duxbury Press; 2006.
10. Hirschmann MT, Konala P, Amsler F, Iranpour F, Friederich NF, Cobb JP. The position and orientation of total knee replacement components: a comparison of conventional radiographs, transverse 2D-CT slices and 3D-CT reconstruction. *J Bone Joint Surg Br* 2011;93:629-633.
11. Kim CH, Chung CK, Hong HS, Kim EH, Kim MJ, Park BJ. Validation of a simple computerized tool for measuring spinal and pelvic parameters. *J Neurosurg Spine* 2012;16:154-162.
12. Donner A, Zou G. Testing the equality of dependent intraclass correlation coefficients. *J R Stat Soc Ser D Stat* 2002;51:367-379.
13. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307-310.
14. Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. *Ultrasound Obstet Gynecol* 2003;22:85-93.
15. Johnsson AA, Fagman E, Vikgren J, Fisichella VA, Boijsen M, Flinck A, et al. Pulmonary nodule size evaluation with chest tomosynthesis. *Radiology* 2012;265:273-282.
16. Bland M. Correction to section "Measuring agreement using repeated measurements" in Bland and Altman (1986) [Internet]. 2009 July 3 [cited 2016 Dec 19]. Available from: https://www-users.york.ac.uk/~mb55/meas/repeated.htm.
17. Hanneman SK. Design, analysis, and interpretation of method-comparison studies. *AACN Adv Crit Care* 2008;19:223-234.
18. Bruton A, Conway JH, Holgate ST. Reliability: what is it, and how is it measured? *Physiotherapy* 2000;86:94-99.
19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-174.
20. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York, NY: John Wiley and Sons; 1981.
21. Altman DG. Practical statistics for medical research. London, UK: Chapman & Hall/CRC; 1991.
22. Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. 3rd ed. Hoboken, NJ: John Wiley & Sons; 2003.
23. StataCorp. STATA base reference manual (release 13). College Station, TX: Stata Press; 2013.