Research Series

Understanding Interobserver Agreement: The Kappa Statistic

Anthony J. Viera, MD; Joanne M. Garrett, PhD

Items such as physical exam findings, radiographic interpretations, or other diagnostic tests often rely on some degree of subjective interpretation by observers. Studies that measure the agreement between two or more observers should include a statistic that takes into account the fact that observers will sometimes agree or disagree simply by chance. The kappa statistic (or kappa coefficient) is the most commonly used statistic for this purpose. A kappa of 1 indicates perfect agreement, whereas a kappa of 0 indicates agreement equivalent to chance. A limitation of kappa is that it is affected by the prevalence of the finding under observation. Methods to overcome this limitation have been described.

(Fam Med 2005;37(5):360-3.)

In reading medical literature on diagnosis and interpretation of diagnostic tests, our attention is generally focused on items such as sensitivity, specificity, predictive values, and likelihood ratios. These items address the validity of the test. But if the people who actually interpret the test cannot agree on the interpretation, the test results will be of little use.

Let us suppose that you are preparing to give a lecture on community-acquired pneumonia. As you prepare for the lecture, you read an article titled, "Diagnosing Pneumonia by History and Physical Examination," published in the *Journal of the American Medical Association* in 1997.[1] You come across a table in the article that shows agreement on physical examination findings of the chest. You see that there was 79% agreement on the presence of wheezing with a kappa of 0.51 and 85% agreement on the presence of tactile fremitus with a kappa of 0.01. How do you interpret these levels of agreement taking into account the kappa statistic?

Accuracy Versus Precision

When assessing the ability of a test (radiograph, physical finding, etc) to be helpful to clinicians, it is important that its interpretation is not a product of guesswork. This concept is often referred to as *precision* (though some incorrectly use the term *accuracy*). Recall the analogy of a target and how close we get to the bull's-eye (Figure 1). If we actually hit the bull's-eye (representing agreement with the gold standard), we are accurate. If all our shots land together, we have good precision (good reliability). If all our shots land together and we hit the bull's-eye, we are accurate as well as precise.

It is possible, however, to hit the bull's-eye purely by chance. Referring to Figure 1, only the center black dot in target A is accurate, and there is little precision (poor reliability about where the shots land). In B, there is precision but not accuracy. C demonstrates neither accuracy nor precision. In D, the black dots are both accurate and precise. The lack of precision in A and C could be due to chance, in which case, the bull's-eye shot in A was just "lucky." In B and D, the groupings are unlikely due to chance.

Precision, as it pertains to agreement between observers (interobserver agreement), is often reported as a kappa statistic.[2] Kappa is intended to give the reader a quantitative measure of the magnitude of agreement between observers. It applies not only to tests such as radiographs but also to items like physical exam findings, eg, presence of wheezes on lung examination as noted earlier. Comparing the presence of wheezes on lung examination to the presence of an infiltrate on a chest radiograph assesses the validity of the exam finding to diagnose pneumonia. Assessing whether the examiners agree on the presence or absence of wheezes (regardless of validity) assesses precision (reliability).

From the Robert Wood Johnson Clinical Scholars Program, University of North Carolina.

Figure 1

Accuracy and Precision

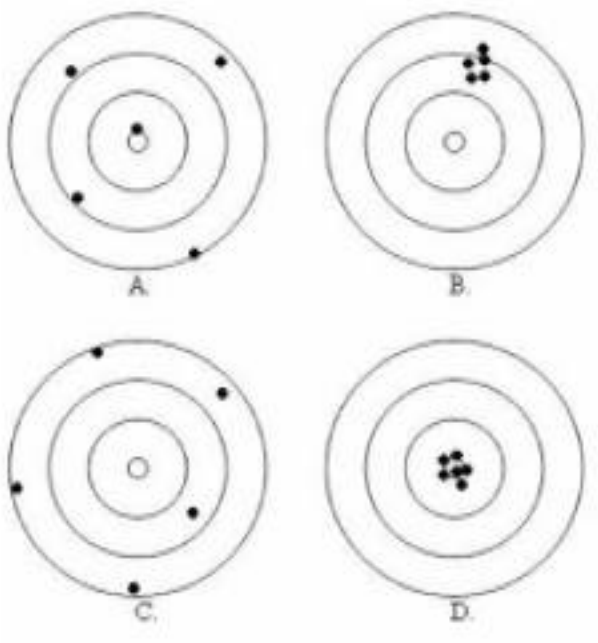

Table 1

Interobserver Variation

Usefulness of Noon Lectures

		Resident 1—Lectures Helpful?		
		Yes	*No*	*Total*
Resident 2—Lectures Helpful?	*Yes*	**15**	5	20
	No	10	**70**	80
	Total	25	75	100

Data Layout

		Observer 1—Result		
		Yes	*No*	*Total*
Observer 2—Result	*Yes*	**a**	b	m_1
	No	c	**d**	m_0
	Total	n_1	n_0	n

(a) and (d) represent the number of times the two observers agree while (b) and (c) represent the number of times the two observers disagree. If there are no disagreements, (b) and (c) would be zero, and the observed agreement (p_o) is 1, or 100%. If there are no agreements, (a) and (d) would be zero, and the observed agreement (p_o) is 0.

Calculations:
Expected agreement

$$p_e = [(n_1/n) * (m_1/n)] + [(n_o/n) * (m_o/n)]$$

In this example, the expected agreement is:
$p_e = [(20/100) * (25/100)] + [(75/100) * (80/100)] = 0.05 + 0.60 = 0.65$

Kappa, K

$$= \frac{(p_o - p_e)}{(1 - p_e)} = \frac{0.85 - 0.65}{1 - 0.65} = 0.57$$

The Kappa Statistic

Interobserver variation can be measured in any situation in which two or more independent observers are evaluating the same thing. For example, let us imagine a study in which two family medicine residents are evaluating the usefulness of a series of 100 noon lectures. Resident 1 and Resident 2 agree that the lectures are useful 15% of the time and not useful 70% of the time (Table 1). If the two residents randomly assign their ratings, however, they would sometimes agree just by chance. Kappa gives us a numerical rating of the degree to which this occurs.

The calculation is based on the difference between how much agreement is actually present ("observed" agreement) compared to how much agreement would be expected to be present by chance alone ("expected" agreement). The data layout is shown in Table 1. The observed agreement is simply the percentage of all lectures for which the two residents' evaluations agree, which is the sum of a + d divided by the total n in Table 1. In our example, this is 15+70/100 or 0.85.

We may also want to know how different the observed agreement (0.85) is from the expected agreement (0.65). Kappa is a measure of this difference, standardized to lie on a -1 to 1 scale, where 1 is perfect agreement, 0 is exactly what would be expected by chance, and negative values indicate agreement less than chance, ie, potential systematic disagreement between the observers. In this example, the kappa is 0.57. (For calculations, see Table 1.)

Interpretation of Kappa

What does a specific kappa value mean? We can use the value of 0.57 from the example above. Not everyone would agree about whether 0.57 constitutes "good" agreement. However, a commonly cited scale is represented in Table 2.[3] It turns out that, using this scale, a kappa of 0.57 is in the "moderate" agreement range between our two observers. Remember that perfect agreement would equate to a kappa of 1, and chance agreement would equate to 0. Table 2 may help you "visualize" the interpretation of kappa. So, residents in this hypothetical study seem to be in moderate agreement that noon lectures are not that helpful.

When interpreting kappa, it is also important to keep in mind that the estimated kappa itself could be due to chance. To report a *P* value of a kappa requires calcula-

Table 2

Interpretation of Kappa

	Poor	Slight	Fair	Moderate	Substantial	Almost perfect
Kappa	0.0	.20	.40	.60	.80	1.0

Kappa	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

tion of the variance of kappa and deriving a z statistic, which are beyond the scope of this article. A confidence interval for kappa, which may be even more informative, can also be calculated. Fortunately, computer programs are able to calculate kappa as well as the *P* value or confidence interval of kappa at the stroke of a few keys. Remember, though, the *P* value in this case tests whether the estimated kappa is not due to chance. It does not test the strength of agreement. Also, *P* values and confidence intervals are sensitive to sample size, and with a large enough sample size, any kappa above 0 will become statistically significant.

Weighted Kappa

Sometimes, we are more interested in the agreement across major categories in which there is meaningful difference. For example, let's suppose we had five categories of "helpfulness of noon lectures:" "very helpful," "somewhat helpful," "neutral," "somewhat a waste," and "complete waste." In this case, we may not care whether one resident categorizes as "very helpful" while another categorizes as "somewhat helpful," but we might care if one resident categorizes as "very helpful" while another categorizes as "complete waste." Using a clinical example, we may not care whether one radiologist categorizes a mammogram finding as normal and another categorizes it as benign, but we do care if one categorizes it as normal and the other as cancer.

A weighted kappa, which assigns less weight to agreement as categories are further apart, would be reported in such instances.[4] In our previous example, a disagreement of normal versus benign would still be credited with partial agreement, but a disagreement of normal versus cancer would be counted as no agreement. The determination of weights for a weighted kappa is a subjective issue on which even experts might disagree in a particular setting.

Table 3

Usefulness of Noon Lectures, With Low Prevalence of Helpful Lectures

		Resident 1— Lectures Helpful?		
		Yes	*No*	*Total*
Resident 2— Lectures Helpful?	*Yes*	**1**	6	7
	No	9	**84**	93
	Total	10	90	100

Calculations:

Observed agreement, $p_o = \frac{1+84}{100} = 0.85$

Expected agreement, p_e =[(7/100) * (10/100)] + [(93/100) * (90/100)] = 0.007 + .837 = 0.844

Calculating kappa:

$K = \frac{(p_o - p_e)}{(1-p_e)} = \frac{0.85-0.844}{1-0.844} = 0.04$

K = .038

Scott's pi = .036

Krippendorff's alpha =- .041

A Paradox

Returning to our original example on chest findings in pneumonia, the agreement on the presence of tactile fremitus was high (85%), but the kappa of 0.01 would seem to indicate that this agreement is really very poor. The reason for the discrepancy between the unadjusted level of agreement and kappa is that tactile fremitus is such a rare finding, illustrating that kappa may not be reliable for rare observations. Kappa is affected by prevalence of the finding under consideration much like predictive values are affected by the prevalence of the disease under consideration.[5] For rare findings, very low values of kappa may not necessarily reflect low rates of overall agreement.

Returning for a moment to our hypothetical study of the usefulness of noon lectures, let us imagine that the prevalence of a truly helpful noon lecture is very low, but the residents know it when they experience it. Likewise, they know (and will say) that most others are not helpful. The data layout might look like Table 3. The observed agreement is high at 85%. However, the kappa (calculation shown in Table 3) is low at .04, suggesting only poor to slight agreement when accounting for chance. One method to account for this paradox, put simply, is to distinguish between agreement on the two levels of the finding (eg, agreement on positive ratings compared to agreement on negative ratings). Feinstein and Cicchetti have published detailed papers on this paradox and methods to resolve it.[5,6] For now, understanding of kappa and recognizing this important limitation will allow the reader to better analyze articles reporting interobserver agreement.

As the values for Scott's pi and Krippendorff's alpha show, the problem identified here for kappa also holds for pi and alpha. Also for ICC (agreement, dummy coding 0,1) = single measure = .041 and average measure = .078.

Conclusions

This article sought to provide a basic overview of kappa as one measure of interobserver agreement. There are other methods of assessing interobserver agreement, but kappa is the most commonly reported measure in the medical literature. Kappa makes no distinction among various types and sources of disagreement. Because it is affected by prevalence, it may not be appropriate to compare kappa between different studies or populations. Nonetheless, kappa can provide more information than a simple calculation of the raw proportion of agreement.

Acknowledgment: Dr Viera is currently funded as a Robert Wood Johnson Clinical Scholar.

Corresponding Author: Address correspondence to Dr Viera, University of North Carolina, Robert Wood Johnson Clinical Scholars Program, CB 7105, 5034 Old Clinic Building, Chapel Hill, NC 27599-7105. 919-966-3712. Fax: 919-843-9237. anthony_viera@med.unc.edu.

References

1. Metlay JP, Wishwa NK, Fine MJ. Does this patient have community-acquired pneumonia? Diagnosing pneumonia by history and physical examination. JAMA 1997;278:1440-5.
2. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960;20:37-46.
3. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
4. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 1968;70:213-20.
5. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543-9.
6. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43:551-8.