Measures of Interrater Agreement

Jayawant N. Mandrekar, PhD

Abstract: Kappa statistics is used for the assessment of agreement between two or more raters when the measurement scale is categorical. In this short summary, we discuss and interpret the key features of the kappa statistics, the impact of prevalence on the kappa statistics, and its utility in clinical research. We also introduce the weighted kappa when the outcome is ordinal and the intraclass correlation to assess agreement in an event the data are measured on a continuous scale.

Key Words: Agreement, Kappa, Weighted kappa, Intraclass correlation.

(*J Thorac Oncol.* 2011;6: 6–7)

In most clinical studies, the goal is to assess the associations between the variables of interest. Nevertheless, in the case of reliability studies, the primary interest is to enumerate the reproducibility of the same variable. For example, the diagnosis of benign versus malignant nodules as rated by two radiologists or the rating of symptom experiences by patients with cancer and caregivers. The application of kappa statistics is only appropriate in cases where agreement between the two raters is of primary interest. If one of the ratings is considered as a gold standard, then the appropriate measure is sensitivity and/or specificity.[1] In this review, we introduce the kappa statistic, outline its utility in clinical research, and discuss the impact of prevalence on the kappa estimates. We also introduce the weighted kappa statistics when the outcome is ordinal and the intraclass correlation when data are measured on a continuous scale.

KAPPA

Dajczman et al.[2] compared patient-rated performance status score with physician-rated Eastern Cooperative Oncology Group performance scale among patients with advanced non-small cell lung cancer (NSCLC). The effect of each rating on the eligibility of the patient for a hypothetical clinical trial was assessed. Moderate level of agreement was reported using the kappa statistic (0.42). Authors also presented results as percent agreement (0.72). To explain how these statistics are computed, consider an example where two radiologists independently review and classify images as benign or malignant from 100 patients using diagnostic modalities (x-ray or positron emission tomography [PET] or computer tomographic [CT] scans) (Table 1).

The observed agreement (P_0) between the raters is given by $(30 + 50)/100 = 0.80$. Nevertheless, this includes the expected agreement, which is the agreement by chance alone (P_e) and the agreement beyond chance. Assuming the rating from the two radiologists is independent of one another, we would have expected $(16 + 36)/100 = 0.52$ agreement by chance alone. This is computed by estimating the expected frequencies in each cell using a combination of row and column totals; in this case, the expected number of malignant nodules called by both readers is 16 (i.e., $40 \times 40/100$), and the expected number of benign nodules called by both readers is 36 (i.e., $60 \times 60/100$). This implies that the agreement beyond chance would have been $P_0 - P_e$, which is $0.80 - 0.52 = 0.38$. The maximum value for P_0 is 1, which happens when the ratings from the two radiologists agree on each of the 100 diagnoses. This implies that the maximum value for $P_0 - P_e$ is $1 - P_e$. Because of the limitation of the simple proportion of agreement and to keep the maximum value of the proposed measure to be 1, Cohen[3] proposed kappa as a measure of interrater agreement. It is calculated as follows: $(P_0 - P_e)/(1 - P_e)$, where $(1 - P_e)$ is interpreted as the proportion of the cases for which the hypothesis of no association would predict disagreement between the raters. Kappa values below 0.4 represent poor agreement, values between 0.4 and 0.75 indicate fair to good agreement, and values of 0.75 and higher represent excellent agreement.[4] In this example, the kappa statistic would be $(0.80 - 0.52)/(1 - 0.52) = 0.58$, which indicates fair to good agreement between the raters.

IMPACT OF PREVALENCE ON KAPPA STATISTICS

Kappa statistics is dependent on the prevalence of the disease. Returning to the example in Table 1, keeping the proportion of observed agreement at 80%, and changing the prevalence of malignant cases to 85% instead of 40% (i.e., higher disease prevalence), the proportion of expected agreement is 0.745 (Table 2). Thus, the kappa statistics becomes $(0.80 - 0.745)/(1 - 0.745) = 0.22$, instead of 0.58. Similarly, if the disease prevalence is low, i.e., higher proportion of the cases is benign, the kappa statistics would be smaller compared with Table 1. Thus, for two studies with the

Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota.

Disclosure: The author declares no conflicts of interest.

Address for correspondence: Jayawant N. Mandrekar, PhD, Department of Health Sciences Research, Mayo Clinic, 200 1st ST SW, Rochester, MN 55905. E-mail: mandrekar.jay@mayo.edu

Copyright © 2010 by the International Association for the Study of Lung Cancer

ISSN: 1556-0864/11/0601-0006

Copyright © 2010 by the International Association for the Study of Lung Cancer.

TABLE 1. Benign vs. Malignant Ratings of Two Independent Raters

	Rater 1		
Rater 2	**Malignant**	**Benign**	**Total**
Malignant	30	10	40
Benign	10	50	60
Total	40	60	100

TABLE 2. Benign vs. Malignant Ratings of Two Independent Raters, When the Disease Prevalence is High

	Rater 1		
Rater 2	**Malignant**	**Benign**	**Total**
Malignant	75	10	85
Benign	10	5	15
Total	85	15	100

same proportion of observed agreement, the maximum value for the kappa statistics would occur in the study where the prevalence is closer to 50%. In other words, in extreme cases where prevalence is very low (e.g., screening population) or very high (i.e., diagnostic or research studies where sample is enriched with positive cases), the kappa statistic would be smaller compared with Table 1. Thus, one needs to exercise caution while comparing kappa values from different studies.

WEIGHTED KAPPA AND INTRACLASS CORRELATION

The simple kappa statistics discussed earlier in the text is relevant to binary ratings. There are several scenarios where ratings could be given on more than two categories. For example, consider stage of the disease, which is ordinal, where the jumps between two consecutive staging levels, stages 1 to 2 versus stages 2 to 3, are not equal. In this case, a modified version of the kappa statistics called as the weighted kappa is calculated allowing one to assign different weights to the different levels. Weighted kappa is the same as simple kappa when there are only two ordered categories. A study by Mac Manus et al.[5] prospectively compared the prognostic value of early posttreatment PET and CT scanning in a cohort of patients with NSCLC treated with radical radiotherapy. Agreement between PET and CT was assessed using weighted kappa, which showed poor agreement between the two modalities (weighted kappa = 0.35).

In certain clinical studies, agreement between the raters is assessed for a clinical outcome that is measured on a continuous scale. In such instances, intraclass correlation is calculated as a measure of agreement between the raters. Intraclass correlation is equivalent to weighted kappa under certain conditions, see the study by Fleiss and Cohen[6,7] for details. In a study by Tyng et al.,[8] Intraclass correlation (ICC) was used to compare the gross volume of lung tumors of patients with NSCLC as defined by specialized radiologist and radiotherapists of a cancer center. The findings suggested an excellent agreement between the two (ICC = 0.94 and 95% confidence interval: 0.87–98).

SUMMARY

Studies designed to quantify the agreement between the raters can be analyzed using kappa statistic, weighted kappa, or intraclass correlation. It is not uncommon to present results from kappa, weighted kappa, and ICC in a single study when measurements are taken on a continuous scale. For example, a study by Guyatt et al.[9] reported the reliability of the assessment of mediastinal lymph node size in CT scan of the thorax. Authors presented results using each of the three approaches; kappa, weighted kappa, and ICC. Kappa was calculated using the node size as enlarged ($\geq$1 cm) versus not enlarged ($<$1 cm), weighted kappa was calculated using creating a variable with four categories ($<$1 cm, 1–1.5 cm, 1.5–2 cm, and $\geq$2 cm), and intraclass correlation was calculated using the actual size measurements. These statistics provide a useful tool in the quantification of the agreement between raters.

REFERENCES

1. Mandrekar JN. Simple statistical measures for diagnostic accuracy assessment. *J Thorac Oncol* 2010;5:763–764.
2. Dajczman E, Kasymjanova G, Kreisman H, et al. Should patient-rated performance status affect treatment decisions in advanced lung cancer? *J Thorac Oncol* 2008;3:1133–1136.
3. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960;20:37–46.
4. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–174.
5. Mac Manus MP, Hicks RJ, Matthews JP, et al. Positron emission tomography is superior to computed tomography scanning for response-assessment after radical radiotherapy or chemoradiotherapy in patients with non-small-cell lung cancer. *J Clin Oncol* 2003;21:1285–1292.
6. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas* 1973;33:613–619.
7. Fleiss JL, Cohen J, Everitt BS. Large-sample standard errors of kappa and weighted kappa. *Psychol Bull* 1969;72:323–327.
8. Tyng CJ, Chojniak R, Pinto PN, et al. Conformal radiotherapy for lung cancer: interobservers' variability in the definition of gross tumor volume between radiologists and radiotherapists. *Radiat Oncol* 2009;4:28.
9. Guyatt GH, Lefcoe M, Walter S, et al. Interobserver variation in the computed tomographic evaluation of mediastinal lymph node size in patients with potentially resectable lung cancer. *Chest* 1995;107:116–119.

Copyright © 2010 by the International Association for the Study of Lung Cancer

Copyright © 2010 by the International Association for the Study of Lung Cancer.