Indian Journal of Medical Sciences
(INCORPORATING THE MEDICAL BULLETIN)
VOLUME 62 JUNE 2008 NUMBER 6

ORIGINAL CONTRIBUTIONS

THE RELIABILITY AND DISTINGUISHABILITY OF ULTRASOUND DIAGNOSIS OF OVARIAN MASSES

ALIREZA AKBARZADEH BAGHEBAN, FARID ZAYERI, FATEMEH BARADARAN ANARAKI[1], ZAHRA ELAHIPANAH[1]

ABSTRACT

BACKGROUND: For any radiologist, intra-observer agreement in observing and decision making in diagnosis of any disease is of great importance, and so is observing and reading ultrasound pictures of ovarian masses and distinguishing amongst their categories. AIMS: In this study, the reliability and consistency of ultrasound diagnosis of ovarian tumors have been evaluated. SETTINGS AND DESIGN: Two experienced and three less experienced radiologists assessed ultrasounds of 40 patients of Mirza Koochak Khan Hospital in Tehran, Iran, in 2005. MATERIALS AND METHODS: In this prospective observational study, the ultrasounds were performed by an expert radiologist, with a single apparatus. These ultrasounds have been evaluated separately and independently in two periods (with a 1-week interval). STATISTICAL ANALYSIS USED: Weighted kappa was used to calculate intra-observer agreement (reliability), and two statistical models were applied to assess category distinguishability (consistency). SPSS version 10, SAS version 8, and EXCEL 2003 have been used to do an appropriate statistical analysis. RESULTS: Mean of weighted kappa was 0.81, and mean of distinguishability was 0.995 for our experienced radiologists, due to their superior results. Because of weaker results obtained by the less experienced radiologists, mean of weighted kappa and mean of distinguishability were 0.65 and 0.967 respectively. Overall mean of distinguishability for benign and borderline categories was 0.969; and for malignant and borderline categories, it was 0.987. CONCLUSION: Although experienced radiologists functioned better than the less experienced radiologists, all of them showed appropriate distinguishability and intra-observer agreement in diagnosis and categorization of the ovarian masses. Distinguishing benign category from borderline was more difficult than distinguishing malignant category from borderline. In general, experienced radiologists showed better results compared to less experienced radiologists.

Key words: Distinguishability, ovarian mass, reliability, ultrasound

Department of Biostatistics, Shahid Beheshti University, MC, Tehran, Iran, [1]Department of Radiology, Tehran University of Medical Sciences, Tehran, Iran

Correspondence:
Alireza Akbarzadeh Bagheban,
Biostatistics Department, School of Paramedics,
Shahid Beheshti University, M.C., Darband St.,
Qods Sq. (Tajrish), Tehran, P.O. Box: 19395-4618, Iran
E-mail: akbarzad@sbmu.ac.ir

INTRODUCTION

Suppose a radiologist classifies each ultrasound in a sample on an ordinal scale at two different times, so that the first evaluation has no effect on the second one; we could show these two ratings by a contingency table and assess two important issues:

- Intra-observer agreement of the observer at two different times. This actually is the reliability of the observer in decision making.[1]
- Distinguishability by the observer in categorizing the samples. When we have ordinal categories, distinguishability of these categories is of great concern, which could show us the ability of the observer in differentiating different categories from each other.[2]

The majority of ordered categories are subjective definitions, and distinguishability by an observer between two close categories is difficult, even for those who are experts.[1] In general, to assess reliability and consistency, kappa and weighted kappa coefficients were used.[3-5] Utilizing these by themselves has some disadvantages, and the results could show some errors as well; therefore, many researchers have recommended using statistical models, in addition to measuring these coefficients for arriving at more complete conclusions.[2,6-9] In this study, we have evaluated the first issue by weighted kappa and the second one by statistical models for ovarian mass data.

MATERIALS AND METHODS

This is a prospective observational study. The data were gathered from the radiology department of Mirza Koochak Khan Hospital in Tehran, Iran, in January 2005. After obtaining consent from 40 women whose ultrasounds were performed by an expert radiologist and just with a single apparatus (in order to minimize the performer bias), two experienced radiologists and three less experienced radiologists evaluated these ultrasounds separately and independently and scored them 1 through 3 for benign, borderline, and malignant cases respectively. In a single blind study, each one of these ultrasounds was reevaluated by our observers for a second time after a week. This period (a week) seems reasonable, because our observers would not recall the ultrasounds after a week and we would not encounter loss of quality of ultrasounds in this short period. Cross classification of these observers at two different times provided five different 3×3 tables, and the tables were used as the basis of our analysis.

In this study, intra-observer agreement of the raters has been evaluated by *weighted kappa* (as index of reliability), and the distinguishability

Table 1: Weighted kappa and distinguishability in different 3×3 tables of the observers

Observer	Distinguishability		
	Weighted Kappa	Benign from borderline	Borderline from malignant
Less experienced radiologist 1	0.61	0.9423	0.9717
Less experienced radiologist 2	0.69	0.9601	0.9864
Less experienced radiologist 3	0.65	0.9640	0.9764
Experienced radiologist 1	0.75	0.9867	0.9999
Experienced radiologist 2	0.87	0.9940	0.9999

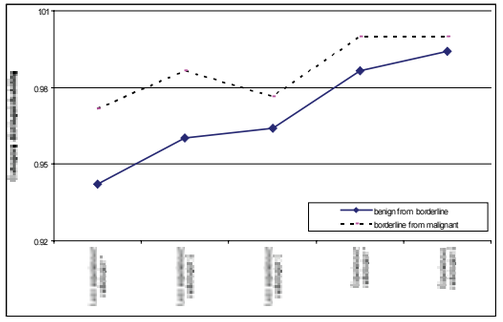

Figure 1: Distinguishability of adjacent categories

by the observers in differentiating categories of ovarian tumor has been assessed by utilizing two statistical models (*'square scores association model'* and *'agreement plus square scores association model'*). These models are special cases of the *'uniform association model'*[10] and the *'Agreement Plus Uniform Association Model'*[9] respectively.

The observers evaluated and reevaluated (after 1 week) 40 different ultrasounds of ovarian masses, separately and independently. Required sample size for these studies (validity and reliability) is usually 15 to 20 cases for quantitative variables and a little more for qualitative variables, so it seemed that 40 cases were enough to achieve our goal and perform our study appropriately.[11]

Distinguishability by the observers in differentiating two adjacent categories could show their ability to determine and diagnose the category or the status of the ovarian mass in ultrasonography.[12,13] The range of this parameter is similar to R^2 coefficient in a regression model and its value varies between zero and one, in which with greater distinguishability by the observer, the value will be closer to one and vice versa.

SPSS version 10 was used for data entry and obtaining appropriate two-dimensional tables. In addition, SAS version 8 was utilized to measure weighted kappa, fit the models, and estimate the models' parameters. To calculate distinguishability and make a figure, we used EXCEL 2003 software.

RESULTS

In this study, we considered three different categories of ovarian mass, and each of the observers classified the ultrasounds at two separate times, so we had five 3×3 tables. The 'square scores association model' had the best fit for the experienced radiologists, and the 'agreement plus square scores association model' had the best fit for the less experienced radiologists.

The experienced radiologists demonstrated high distinguishability in categorizing different categories (minimum 0.98 for benign and borderline [1 and 2] and minimum 0.99 for borderline and malignant [2 and 3] entities),

and there was no significant difference between these two categorization abilities of experienced radiologists. The overall mean of distinguishability for these raters was 0.995, and the mean of weighted kappa for them was 0.81 [Table 1].

The less experienced radiologists demonstrated lower distinguishability in categorizing different categories (minimum 0.95 for benign and borderline [1 and 2] and minimum 0.97 for borderline and malignant [2 and 3] entities) [Figure 1]. These raters had an overall distinguishability mean of 0.967, and it was a little lower compared to the experienced radiologists. Mean of weighted kappa for them was 0.65.

The mean of distinguishability for benign and borderline categories was 0.990 for the experienced radiologists and 0.955 for the less experienced radiologists. Besides, the experienced radiologists and the less experienced radiologists had a mean of 0.999 and 0.978 respectively for distinguishing the borderline and malignant cases.

DISCUSSION

To compare distinguishability demonstrated by the observers in categorizing the samples and assessing intra-observer agreement for each one of them, we computed weighted kappa at first. Although there was no complete intra-observer agreement for these observers at two different times, by considering 0.71 for mean of weighted kappa, it can be stated that there was good overall reliability.[14] Besides, minimum and maximum of weighted kappa in our study have been obtained to be 0.61 and 0.86 respectively.

Our findings confirm the results reported by Amer *et al.*[15] They found 69.4% for the mean intra-observer agreement (kappa = 0.54). One reason for a small difference in reliability index is that they used kappa instead of weighted kappa.

Although the less experienced radiologists demonstrated a lower distinguishability compared to the experienced radiologists, yet this difference was not remarkable; because all the observers had a minimum 0.90 to distinguish between adjacent categories. But for all observers, distinguishability between categories 1 and 2 was lower than that between categories 2 and 3; and experienced radiologists showed better results than the less experienced radiologists.

Generally, for assessing validity and reliability of diagnosing among different categories of ovarian cysts, kappa and weighted kappa coefficients are used.[15] These coefficients show intra-observer agreement generally; and by considering several deficiencies that were reported for them in multiple studies[2,5,7,8] and their inability to show distinguishability by observers, we used statistical models to consider distinguishability demonstrated by them to classify different ordered categories. We could use these results for better future training of raters in big epidemiological studies.

REFERENCES

1. Agresti A. A model for agreement between ratings on an ordinal scale. Biometrics 1988;44:539-48.
2. Perkins SM, Becker MP. Assessing rater

agreement using marginal association models. Stat Med 2002;21:1743-60.
3. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measures 1960;20:37-46.
4. Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213-20.
5. Kraemer HC, Periakoil VS, Noda A. Tutorial in biostatistics, kappa coefficients in medical research. Stat Med 2002;21:2109-29.
6. Koch GG, Landis JR, Freeman JL, Freeman DH, Lehnen RG. A general methodology for the analysis of experiments with repeated measurement of categorical data. Biometrics 1977;33:133-58.
7. Tanner MA, Young MA. Modeling agreement among raters. JASA 1985;80:175-80.
8. Feinstein AR, Cicchetti DV. High agreement but low kappa: I, The problem of two paradoxes. J Clin Epidemiol 1990;43:543-9.
9. May SM. Modeling observer agreement: An alternative to kappa. J Clin Epidemiol 1994;44:1315-24.
10. Goodman LA. Simple models for the analysis of association in cross-classifications having ordered categories. JASA 1979;74:537-52.
11. Fleiss JL. The design and analysis of clinical experiments. 151 ed. New York: John Wiley and Sons; 1999. p. 8.
12. Darroch JN, McCloud PI. Category distinguishability and observer agreement. Aust J Stat 1986;28: 371-88.
13. Becker MP, Agresti A. Log-linear modeling of pairwise interobserver agreement on a categorical scale. Stat Med 1992;11:101-14.
14. Altman DG. Practical statistics for medical research. London England: Chapman and Hall; 1991. p. 404.
15. Amer S, Li TC, Bygrave C, Sprigg A, Saravelos H, Cooke ID. An evaluation of the inter-observer and intra-observer variability of the ultrasound diagnosis of polycystic ovaries. Hum Reprod 2002;17:1616-22.

Source of Support: Nil
Conflict of Interest: None declared.