Radiology

Harold L. Kundel, MD
Marcia Polansky, ScD

Index terms:
Diagnostic radiology, observer performance
Statistical analysis

Published online before print
10.1148/radiol.2282011860
Radiology 2003; 228:303–308

[1] From the Department of Radiology (H.L.K.) and MCP Hahnemann School of Public Health (M.P.), University of Pennsylvania Medical Center, 3600 Market St, Suite 370, Philadelphia, PA 19104. Received November 21, 2001; revision requested January 29, 2002; revision received March 4; accepted March 11. Supported by grant P01 CA 53141 from the National Cancer Institute, National Institutes of Health, U.S. Public Health Service, Department of Health and Human Services. **Address correspondence to** H.L.K. (e-mail: *kundel@rad.upenn.edu*).

Measurement of Observer Agreement[1]

Statistical measures are described that are used in diagnostic imaging for expressing observer agreement in regard to categorical data. The measures are used to characterize the reliability of imaging methods and the reproducibility of disease classifications and, occasionally with great care, as the surrogate for accuracy. The review concentrates on the chance-corrected indices, κ and weighted κ. Examples from the imaging literature illustrate the method of calculation and the effects of both disease prevalence and the number of rating categories. Other measures of agreement that are used less frequently, including multiple-rater κ, are referenced and described briefly.

The statistical analysis of observer agreement in imaging is generally performed for three reasons. First, observer agreement provides information about the reliability of imaging diagnosis. A reliable method should produce good agreement when used by knowledgeable observers. Second, observer agreement can be used to check the consistency of a method for classification of an abnormality that indicates the extent or severity of disease (1) and to determine the reliability of various signs of disease (2). It can also be used to compare the performance of humans and computers (3). Third, observer agreement can provide a general estimate of the value of an imaging technique when an independent method of proving the diagnosis precludes the measurement of sensitivity and specificity or the more general receiver operating characteristic curve. In many clinical situations, imaging provides the best evidence of abnormality. Furthermore, even if an independent method for obtaining proof exists, it may be difficult to use. For every suspected lesion, a biopsy cannot be performed to obtain a specific tissue diagnosis. As we will demonstrate, currently popular measures of agreement do not necessarily reflect accuracy. However, there are statistical techniques for use of the agreement of multiple expert readers (4) or the agreement of multiple tests (5) to estimate the underlying accuracy of the test.

We illustrate the standard methods for description of agreement in regard to categorical data and point out the advantages and disadvantages of the use of these methods. We refer to some of the less common, although not less important, methods but do not describe them. Then we describe some current developments in methods for use of agreement to estimate accuracy. The discussion is limited to data that can be assigned to categories, such as positive or negative; high, medium, or low; class I–V. Data, such as lesion volume or heart size, that are collected on a continuous scale are more appropriately analyzed with methods of correlation.

MEASUREMENT OF AGREEMENT OF TWO READERS

Consider readings of the same 150 images that are reported as either positive or negative by two readers. The results are shown in Table 1 as joint agreement in a 2×2 format, with the responses of each reader as marginal totals. Three general indices of agreement can be derived from Table 1. The overall proportion of agreement, which we will call p_o, is calculated as follows:

$$p_o = \frac{7 + 121}{150} = 0.85.$$

The proportion is useful for calculations, but the result is usually expressed as a percentage. A p_o of 0.85 indicates that the two readers agree in regard to 85% of their interpretations. If the number of negative readings is large relative to the number of positive readings, the agreement in regard to negative readings will dominate the value of

p_o and may give a false impression of performance. For example, suppose that 90% of the cases are actually negative, and two readers agree about all of the negative interpretations but disagree about the positive interpretations. The overall agreement will be at least 90% and may be greater depending on the number of positive interpretations on which they agree. As an alternative to the overall agreement, the positive and negative agreement can be estimated separately. This will give an indication of the type of decision on which readers disagree. The positive agreement, which we will call p_{pos}, is the number of positive readings that both readers agree on divided by all of the positive readings for both readers. For the data in Table 1, the positive agreement is calculated with the following equation:

$$p_{pos} = \frac{7 + 7}{(10 + 7) + (12 + 7)} = 0.39.$$

The negative agreement, which we will call p_{neg}, can be calculated in a similar way as follows:

$$p_{neg} = \frac{121 + 121}{(10 + 121) + (12 + 121)} = 0.92.$$

In the example given in Table 1, although the two readers agree 85% of the time overall, they only agree on positive interpretations 39% of the time, whereas they agree on negative interpretations 92% of the time. The advantage of calculation of p_{pos} and p_{neg} is that any imbalance in the proportion of positive and negative responses becomes apparent, as in the example. The disadvantage is that CIs cannot be calculated.

TABLE 1
Joint Judgment of Two Readers about Same 150 Images

First Reader	Second Reader: Positive for Disease	Second Reader: Negative for Disease	Total
Positive for disease	7	10	17
Negative for disease	12	121	133
Total	19	131	150

TABLE 2
Guidelines for Strength of Agreement Indicated with κ Values

κ Value	Strength of Agreement beyond Chance
<0	Poor
0–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

Note.—Data are from Landis and Koch (8).

TABLE 3
Joint Judgment of Two Readers about Position of Tubes and Catheters on 100 Portable Chest Images

First Reader	Second Reader: Malpositioned	Second Reader: Correctly Positioned	Total
Malpositioned	3	3	6
Correctly positioned	2	92	94
Total	5	95	100

TABLE 4
Joint Judgment of Two Readers about Presence of Signs of Congestive Heart Failure on 100 Portable Chest Images

First Reader	Second Reader: CHF	Second Reader: No CHF	Total
CHF	20	12	32
No CHF	8	60	68
Total	28	72	100

Note.—CHF = congestive heart failure.

COHEN κ

Some of the observer agreement concerning findings of imaging tests can be caused by chance. For example, chance agreement occurs when the readers know in advance that most of the cases are negative and they adopt a reading strategy of reporting a case as negative whenever they are in doubt. Both will have a large percentage of negative agreements because of prior knowledge of the prevalence of negative cases, not because of information obtained from viewing of the images. An index called κ has been developed as a measure of agreement that is corrected for chance. The κ is calculated by subtracting the proportion of the readings that are expected to agree by chance, which we will call p_e, from the overall agreement, p_o, and dividing the remainder by the number of cases on which agreement is not expected to occur by chance. This is demonstrated in Equation (1) as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e}. \quad (1)$$

Another way to view κ is that if the readers read different images and the readings were paired, some agreement, namely p_o, would be observed. The observed agreement would occur purely by chance. The agreement that is expected to occur by chance, which we shall designate p_e, can be calculated. When the readings of different images are compared, the observed value, namely the p_o, should equal the expected value, p_e, because there is no agreement beyond chance and κ is zero.

The joint agreement that is expected because of chance is calculated for each combination with multiplication of the total responses of each reader contained in the marginal totals of the data table. From Table 1, the agreement expected by chance for the joint positive and joint negative responses is calculated with the following equation:

$$p_e = \left(\frac{17}{150} \cdot \frac{19}{150}\right) + \left(\frac{133}{150} \cdot \frac{131}{150}\right) = 0.79.$$

The value for κ is 0.31, as is calculated with this equation:

$$\kappa = \frac{0.85 - 0.79}{1 - 0.79} = 0.31.$$

The standard error, which we will call SE, of κ for a 2 × 2 table can be estimated with the following equation:

$$SE = \sqrt{\frac{p_o(1 - p_o)}{n(1 - p_e)^2}},$$

$$SE(\kappa) = \sqrt{\frac{0.85(1 - 0.85)}{150(1 - 0.79)^2}} = 0.14. \quad (2)$$

A more accurate and more complicated equation for the standard error of κ can be found in most books about statistics (6,7).

The 95% CIs of κ can be calculated as follows:

$$CI_{95\%} = \kappa \pm 1.96 \cdot SE(\kappa). \quad (3)$$

For example, the 95% CIs are 0.31 − 1.96 × 0.14 = 0.04 and 0.31 + 1.96 × 0.14 = 0.58.

Thus, what is the meaning of a κ of 0.31, together with an overall agreement of 0.85? The calculated value of κ can range from −1.00 to +1.00, but for practical purposes the range from zero to +1.00 is of interest. A κ of zero means that there is no agreement beyond chance, and a κ of

TABLE 5
Indices of Agreement for Readings of Two Radiologists Regarding Portable Chest Images for Position of Tubes and Catheters and Signs of Congestive Heart Failure

Agreement Index	Type of Agreement	Tubes and Catheters	Congestive Heart Failure
p_o	Overall	0.95	0.80
p_{pos}	Positive	0.54	0.67
p_{neg}	Negative	0.97	0.86
p_e	Chance	0.90	0.57
κ	Chance corrected	0.52	0.52

TABLE 6
Comparison of Unweighted and Weighted p_o and κ Calculated by Using Four-, Three-, and Two-Response Categories

	Unweighted		Quadratic Weighting	
Categories	p_o	κ	$p_o(w)$	$\kappa(w)$
Four-response	0.55	0.37	0.93	0.76
Three-response	0.66	0.48	0.92	0.71
Two-response	0.82	0.62	0.82	0.62

Note.—Values were calculated for data from Table A1.

1.00 means that there is perfect agreement. Interpretations of intermediate values are subjective. Table 2 shows the strength of agreement beyond chance for various ranges of κ that were suggested by Landis and Koch (8). The choice of intervals is entirely arbitrary but has become ingrained with frequent usage. The values calculated from Table 1 show that there is good overall agreement (p_o = 0.85) but only fair chance-corrected agreement (κ = 0.31). This paradoxical result is caused by the high prevalence of negative cases. Prevalence effects can lead to situations in which the values of κ do not correspond with intuition (9,10). This is illustrated with the data in Tables 3 and 4 that were extrapolated, with a bit of adjustment to make the numbers come out even, from a data set collected during a study of readings in regard to portable chest images obtained in a medical intensive care unit (11). Table 3 shows the agreement of the reports of two of the readers concerning the position of tubes and catheters. An incorrectly positioned tube or catheter was defined as a positive reading. Table 4 shows the agreement in regard to the reports of the same two readers about the presence of radiographic signs of congestive heart failure. The example was chosen because the actual values of κ for the two diagnoses were very close.

The agreement indices for the two types of readings are shown in Table 5. The overall agreement (95%) for the position of tubes and catheters is very high, but so is the agreement according to chance (90%) calculated from the marginal values in Table 3. This results in a low κ of 0.52, which happens to be the same κ as that for congestive heart failure. The result is not intuitively appealing, because a relatively simple decision such as that about the location of a catheter tip should have a higher index of agreement than a more difficult decision such as that concerning a diagnosis of congestive heart failure. Feinstein and Cicchetti (9) have pointed out the paradox of high overall agreement and low κ, and Cicchetti and Feinstein (10) suggest that when investigators report the results of studies of agreement they should include the three indices of κ, positive agreement, and negative agreement. We agree that this is a useful way of showing agreement data, because it provides more details about where disagreements occur and alerts the reader to the possibility of effects caused by prevalence or prior knowledge.

WEIGHTED κ FOR MULTIPLE CATEGORIES

The κ can be calculated for two readers who report results with multiple categories. As the number of categories increases, the value of κ decreases because there is more room for disagreement with more categories. However, when findings are reported by using a ranked variable, the relative importance of disagreement between categories may not be the same for adjacent categories as it is for distant categories. Two readers who consistently disagree about minimal and moderate categories would have the same value for κ calculated in the usual way as would two readers who consistently disagree about minimal and severe categories. A method for calculation of κ has been developed that allows for differences in the importance of disagreements. The usual approach is to assign weights between 1.00 and zero to each agreement pair, where 1.00 represents perfect agreement and zero represents no agreement. Assignment of weights can be very subjective and can confuse comparison of κ values between studies in which different weights were used. For theoretical reasons, Fleiss (7) suggests assignment of weights as follows:

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2}, \qquad (4)$$

where w represents weight, i is the number of the row, j is the number of the column, and k is the total number of categories. The weighting is called quadratic because of the squared terms. An example of the method for calculation of weighted κ by using four categories is presented in the Appendix. In the example in the Appendix, the categories of absent, minimal, moderate, and severe are used. The weighted and unweighted values for p_o and κ are included in Table 6. The calculations were repeated by collapsing the data for four categories first into three and then into two categories: First, minimal and moderate categories were combined, and then minimal, moderate, and severe categories were combined, and these two combinations would be equivalent to normal and abnormal categories, respectively. Table 6 shows that the value of κ increases as the number of categories is decreased, thus indicating better agreement when the fine distinctions are eliminated. The weighted κ is greater than the unweighted κ when multiple categories are used and is the same as the unweighted κ when only two categories are used. Some investigators prefer to use multiple categories because they are a better reflection of actual clinical decisions, and if sensible weighting can be achieved, the weighted κ may reflect the actual agreement better than does the unweighted κ.

ESTIMATION OF κ FOR MULTIPLE READERS

When multiple readers are used, some authors calculate the values of κ for pairs of readers and then compute an average κ for all possible pairs (12–14). Fleiss (7) describes a method for calculation of a κ index for multiple readers. It has not been used very much in diagnostic imaging, although it has been reported in some studies along with values for weighted κ (15).

ADVANTAGES AND DISADVANTAGES OF THE κ INDEX

κ has the advantage that it is corrected for agreement with statistical chance, and there is an accepted method for computing confidence limits and for statistical testing. The main disadvantage of κ is that the scale is not free of dependence on disease prevalence or the number of rating categories. As a consequence, it is difficult to interpret the meaning of any absolute value of κ, although it is still useful in experiments in which a control for prevalence and for the number of categories is used. The prevalence bias makes it difficult to compare the results of clinical studies where disease prevalence may vary; for example, this may occur in studies about the screening and diagnosis of breast cancer. The disease prevalence should always be reported when κ is used to prevent misunderstanding when one is trying to make generalizations.

RELATIONSHIP BETWEEN AGREEMENT AND ACCURACY

High accuracy implies high agreement, but high agreement does not necessarily imply high accuracy. There is no direct way to infer the accuracy in regard to an image-reading task from reader agreement. Accuracy can only be implied from agreement, with the assumption that when readers agree they must be correct. We frequently make this assumption by seeking a consensus diagnosis or by obtaining a second opinion, but it is not always correct. The κ has been shown to be inconsistent with accuracy as measured by the area under the receiver operating characteristic curve (16) and should not be used as a surrogate for accuracy. Different areas under the receiver operating characteristic curve can have the same κ, and the same areas under the receiver operating characteristic curve can have different κ values. For example, Taplin et al (14) studied the accuracy and agreement of single- and double-reading screening mammograms by using the area under the receiver operating characteristic curve and κ. The study included 31 radiologists who read 120 mammograms. The mean area under the receiver operating characteristic curve for single-reading mammograms was 0.85, and that for double-reading mammograms was 0.87. However, the average unweighted κ for patients with cancer was 0.41 for single-reading mammograms and 0.71 for double-reading mammograms. The average unweighted κ for patients without cancer was 0.26 for single-reading mammograms and 0.34 for double-reading mammograms. Double reading of mammograms resulted in better agreement but not in better accuracy.

If we assume that agreement implies accuracy, then we can use measurements of observed agreement to set a lower limit for accuracy. Suppose two readers agree with respect to interpretation in 50% of the cases; then, by implication, they are both correct with respect to interpretation in 50% of the cases about which they agree and one of them is correct with respect to interpretation in half (25% of the total) of the cases about which they disagree. Therefore, the overall accuracy of the readings is 75%. Typically, in radiology, observed between-reader agreement is 70%–80%, implying an accuracy that is 85%–90% (ie, 70% + 30%/2 to 80% + 20%/2).

Some new approaches to estimation of accuracy from agreement have been proposed. These approaches are based on the assumption that when a majority of readers agree about a diagnosis they are likely to be right (4,17). We have proposed the use of a technique called mixture distribution analysis (4,18). At least five readers report the cases by using either a yes-no response or a rating scale. The agreement of the group of readers about each case is fit to a mathematic model, with the assumption that the sample was drawn from a population that consists of easy normal, easy abnormal, and hard cases. With the computer program, the population that best fits the sample is located, and an overall measure of performance that we call the relative percentage agreement is calculated. We have found that the relative percentage agreement has values similar to those obtained

TABLE A1
Frequency of Responses of Two Readers Who Rated a Disease as Absent, Minimal, Moderate, or Severe

	Reader 1				
Reader 2	Absent	Minimal	Moderate	Severe	Total
Absent	34	10	2	0	46
Minimal	6	8	8	2	24
Moderate	2	5	4	12	23
Severe	0	1	2	14	17
Total	42	24	16	28	110

Note.—The frequencies in Table A1 are converted into proportions in Table A2 by dividing by the total number of cases.

TABLE A2
Proportion of Responses of Two Readers Who Rated a Disease as Absent, Minimal, Moderate, or Severe

	Reader 1				
Reader 2	Absent	Minimal	Moderate	Severe	Total
Absent	0.31	0.09	0.02	0	0.42
Minimal	0.05	0.07	0.07	0.02	0.22
Moderate	0.02	0.05	0.04	0.11	0.21
Severe	0	0.01	0.02	0.13	0.15
Total	0.38	0.22	0.15	0.25*	1.00

* Value was rounded.

TABLE A3
Quadratic Weights for 4 × 4 Table

	Absent, 1	Minimal, 2	Moderate, 3	Severe, 4
Absent, 1	1.0	0.89	0.56	0
Minimal, 2	0.89	1.00	0.89	0.56
Moderate, 3	0.56	0.89	1.00	0.89
Severe, 4	0	0.56	0.89	1.00

Note.—Numbers 1–4 are weighting factors that correspond to the respective category.

TABLE A4
Weighted Proportion of Observed and Expected Responses

Disease Rating Category	Observed Weighted Proportions for Disease Rating Category				Expected Weighted Proportions for Disease Rating Category			
	Absent	Minimal	Moderate	Severe	Absent	Minimal	Moderate	Severe
Absent	0.31	0.08	0.01	0	0.16	0.08	0.03	0
Minimal	0.05	0.07	0.06	0.01	0.07	0.05	0.03	0.03
Moderate	0.01	0.04	0.04	0.10	0.04	0.04	0.03	0.05
Severe	0	0.01	0.02	0.13	0	0.02	0.02	0.04

by using receiver operating characteristic curve analysis with proved cases (18,19).

CONCLUSION

Formal evaluations of imaging technology by using reader agreement started in 1947 with the publication of an article about tuberculosis case finding by using four different chest imaging systems (20). The author of an editorial that accompanied the article expressed surprise that there was so much disagreement (21). History repeated itself when an article about agreement in screening mammography that showed considerable reader variability (22) was published; this article was accompanied by an editorial in which the author expressed surprise in regard to the extent of disagreement (23). The consensus of a group of physicians is frequently the only basis for determination of a difficult diagnostic decision. Studies of pathologists who classify cancer have shown levels of disagreement are similar to those associated with hard decisions in radiology (24). Agreement usually results from informal discussion; however, the method used to obtain agreement can have a large influence on the decision outcome (25). Formal procedures that are used to achieve agreement have been proposed (26); although they can minimize individual bias in achieving a consensus, they are rarely used. We hope that this brief review will stimulate greater use of existing statistics for characterization of agreement and further exploration of new methods.

APPENDIX

Consider a data set in Table A1 that consists of four categories. The frequencies in Table A1 are converted into proportions, which are included in Table A2, by dividing the data by the total number of cases.

Table A3 shows the quadratic weights calculated by using Equation (4), as presented earlier:

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2},$$

where w represents weight, i is the number of the row, j is the number of the column, and k is the total number of categories. It is assumed that disagreement between adjacent categories (ie, disagreement for absent to minimal is 0.89) is not as important as that between distant categories (ie, disagreement for absent to severe is zero).

The weighted observed agreement is calculated by multiplying the proportion of responses in each cell of the 4 × 4 table by the corresponding weighting factor. The calculations for the first row are as follows: 0.31 × 1.00 = 0.31, 0.09 × 0.89 = 0.08, 0.02 × 0.56 = 0.01, and 0 × 0 = 0.

The results for observed weighted proportions are presented in Table A4. The expected agreement is calculated by multiplying the row and column total for each cell of the 4 × 4 table by the corresponding weighting factor. The calculations for the first row are as follows: (0.42 × 0.38) × 1.00 = 0.16, (0.42 × 0.22) × 0.89 = 0.08, (0.42 × 0.15) × 0.56 = 0.03, and (0.42 × 0.25) × 0 = 0.

The results for expected weighted proportions are presented in Table A4. The sum of all of the cells in regard to observed weighted proportions (sum, 0.93) in Table A4 is the weighted observed agreement, which we call $p_o(w)$, and the sum of all of the cells in regard to expected weighted proportions (sum, 0.70) in Table A4 is the weighted expected agreement, which we call $p_e(w)$. When we apply the equation for κ to the weighted values, we get a weighted κ index of 0.76, which is calculated with the following equation:

$$\kappa(w) = \frac{p_o(w) - p_e(w)}{1 - p_e(w)}.$$

An unweighted κ can be calculated by using the sum of the diagonal cells in Table A2, or 0.31 + 0.07 + 0.04 + 0.13 = 0.55, to calculate the observed agreement and the sum of the diagonal cells in Table A4 with regard to expected weighted proportions, or 0.16 + 0.05 + 0.03 + 0.04 = 0.28, to calculate the expected agreement. The unweighted κ is 0.37.

The calculation of the appropriate standard error and the use of the standard error for testing either the hypothesis that κ is different from zero or that κ is different from a value other than zero is beyond the scope of this article but is in most basic statistical texts (6,7).

GLOSSARY

Below is a list of common terms and definitions related to the measurement of observer agreement.

Accuracy.—This value is the likelihood of the interpretation being correct when compared with an independent standard.

Agreement.—This term represents the likelihood that one reader will indicate the same responses as another reader.

Attributes.—An attribute is a categorical variable that represents a property of the object being imaged (eg, tumor descriptors such as mass, calcification, and architectural distortion).

Categorical variables.—Categorical variables are variables that can be assigned to specific categories. Categorical variables can be either ranked variables or attributes.

κ.—The κ value is an overall measure of agreement that is corrected for agreement by chance. It is sensitive to disease prevalence.

Marginal sums.—A marginal sum is the sum of the responses in a single row or column of the data table, and it represents the total response of one of the readers.

Measurement variable.—Measurement variables are variables that can be measured or counted. They are generally divided into continuous variables (eg, lesion diameter or volume) and discrete variables (eg, number

of lesions, expressed as whole numbers but never as decimal fractions).

Prevalence.—Prevalence is the proportion of a particular class of cases in the population being studied.

Ranked variables.—Ranked variables are categorical variables that have a natural order, such as stage of a disease, histologic grade, or discrete severity index (ie, mild, moderate, or severe).

Reliability.—Reliability is the likelihood that one reader will provide the same responses as those provided by a large consensus group.

Weighted κ.—The weighted κ is an overall measure of agreement that is corrected for agreement by chance; a weighting factor is applied to each pair of disagreements to account for the importance of the disagreement.

References

1. Baker JA, Kornguth PJ, Floyd CE. Breast imaging reporting and data system standardized mammography lexicon: observer variability in lesion description. AJR Am J Roentgenol 1996; 166:773–778.
2. Markus JB, Somers S, Franic SE, et al. Interobserver variation in the interpretation of abdominal radiographs. Radiology 1989; 171:69–71.
3. Tiitola M, Kivisaari L, Tervahartiala P, et al. Estimation or quantification of tumour volume? CT study on irregular phantoms. Acta Radiol 2001; 42:101–105.
4. Polansky M. Agreement and accuracy: mixture distribution analysis. In: Beutel J, VanMeter R, Kundel H, eds. Handbook of imaging physics and perception. Bellingham, Wash: Society of Professional Imaging Engineers, 2000; 797–835.
5. Henkelman RM, Kay I, Bronskill MJ. Receiver operating characteristic (ROC) analysis without truth. Med Decis Making 1990; 10:24–29.
6. Agresti A. Categorical data analysis. New York, NY: Wiley, 1990; 366–370.
7. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York, NY: Wiley, 1981; 212–236.
8. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159–174.
9. Feinstein A, Cicchetti D. High agreement but low kappa. I. The problem of two paradoxes. J Clin Epidemiol 1990; 43:543–549.
10. Cicchetti D, Feinstein A. High agreement but low kappa. II. Resolving the paradoxes. J Clin Epidemiol 1990; 43:551–558.
11. Kundel HL, Gefter W, Aronchick J, et al. Relative accuracy of screen-film and computed radiography using hard and soft copy readings: a receiver operating characteristic analysis using bedside chest radiographs in a medical intensive care unit. Radiology 1997; 205:859–863.
12. Epstein DM, Dalinka MK, Kaplan FS, et al. Observer variation in the detection of osteopenia. Skeletal Radiol 1986; 15:347–349.
13. Herman PG, Khan A, Kallman CE, et al. Limited correlation of left ventricular end-diastolic pressure with radiographic assessment of pulmonary hemodynamics. Radiology 1990; 174:721–724.
14. Taplin SH, Rutter CM, Elmore JG, et al. Accuracy of screening mammography using single versus independent double interpretation. AJR Am J Roentgenol 2000; 174:1257–1262.
15. Robinson PJ, Wilson D, Coral A, et al. Variation between experienced observers in the interpretation of accident and emergency radiographs. Br J Radiol 1999; 72:323–330.
16. Swets JA. Indices of discrimination or diagnostic accuracy: their ROCs and implied models. Psychol Bull 1986; 99:100–117.
17. Uebersax JS. Modeling approaches for the analysis of observer agreement. Invest Radiol 1992; 27:738–743.
18. Kundel HL, Polansky M. Mixture distribution and receiver operating characteristic analysis of bedside chest imaging using screen-film and computed radiography. Acad Radiol 1997; 4:1–7.
19. Kundel HL, Polansky M. Comparing observer performance with mixture distribution analysis when there is no external gold standard. In: Kundel HL, ed. Medical imaging 1998: image perception. Bellingham, Wash: Society of Professional Imaging Engineers, 1998; 78–84.
20. Birkelo CC, Chamberlain WE, Phelps PS, et al. Tuberculosis case finding: a comparison of the effectiveness of various roentgenographic and photofluorographic methods. JAMA 1947; 133:359–366.
21. The "personal equation" in the interpretation of a chest roentgenogram (editorial). JAMA 1947; 133:399–400.
22. Elmore JG, Wells CK, Lee CH, et al. Variability in radiologists' interpretation of mammograms. N Engl J Med 1994; 331:1493–1499.
23. Kopans DB. Accuracy of mammographic interpretation (editorial). N Engl J Med 1994; 331:1521–1522.
24. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 1977; 33:363–374.
25. Revesz G, Kundel HL, Bonitatibus M. The effect of verification on the assessment of imaging techniques. Invest Radiol 1983; 18:194–198.
26. Hillman BJ, Hessel SJ, Swensson RG, Herman PG. Improving diagnostic accuracy: a comparison of interactive and Delphi consultations. Invest Radiol 1977; 12:112–115.