J Clin Epidemiol Vol. 41, No. 10, pp. 949–958, 1988
Printed in Great Britain

0895-4356/88 $3.00 + 0.00
Pergamon Press plc

Variance and Dissent

A REAPPRAISAL OF THE KAPPA COEFFICIENT

W. DOUGLAS THOMPSON[1] and STEPHEN D. WALTER[2]

[1]Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College Street, New Haven, CT 06510, U.S.A. and [2]Department of Clinical Epidemiology and Biostatistics, McMaster University, 1200 Main Street West, Hamilton, Ontario, Canada L8N 3Z5

(*Received in revised form* 9 *November* 1988)

Abstract—Kappa is frequently used in epidemiology as an index of the quality of measurement for binary characteristics. The authors discuss the strong dependence of kappa on true prevalence, and they examine the relationship of the value of kappa to the degree of attenuation of the odds ratio that results from non-differential misclassification. It is concluded that under certain circumstances kappa can be interpreted as an indicator of validity, i.e. unbiasedness of the odds ratio, rather than simply as one of reliability. Cautions are stressed regarding (1) possible variation in the quality of measurement and (2) possible lack of independence of errors for the paired measurements from which kappa is calculated. An important implication for the design of reliability studies is that they should be conducted in populations where the distribution of the factor of interest is similar to that for the settings in which the measurement technique will ultimately be applied.

Kappa Validity Reliability Misclassification Odds ratio Attenuation

INTRODUCTION

In epidemiologic research, as in any scientific endeavor, the interpretability of observed results depends to a considerable extent on the accuracy of the measurements made. In epidemiology the measurement process often entails the classification of individuals according to whether given characteristics (disease, exposure, etc.) are present or absent. The purpose of this paper is to explore conceptual and practical issues that arise when the kappa coefficient is employed to quantify the quality of measurement for a binary variable. Some of these issues have received occasional attention in the area of psychiatric research [1–4], but are apparently not well known among epidemiologists. In addition, we propose a new basis for interpreting the kappa coefficient and for judging its usefulness in a variety of situations.

There are two essential features in the accuracy of binary classifications. One is sensitivity (i.e. the proportion of those who truly have the characteristic that are so classified), and the other is specificity (i.e. the proportion of those who truly do not have the characteristic that are so classified). When the actual presence or absence of the characteristic is known for a group of individuals, then sensitivity and specificity for an imperfect classification may be estimated directly. In the more typical situation, however, the "truth" is not known, and in order to judge the quality of measurement one must settle for an assessment of agreement between multiple imperfect sources of information or between multiple measurements using a single imperfect source of information.

The simple percent agreement between two binary classifications has been long recognized as a potentially misleading index because of the often substantial chance agreement that occurs even if there is no systematic tendency for the two measurements to classify the same individuals similarly [5]. A measure that corrects for this chance-expected agreement is the kappa

coefficient as applied to 2×2 tables [6]. Specifically,

$$\text{Kappa} = \frac{\text{Observed} - \text{Expected Proportion in Agreement}}{1 - \text{Expected Proportion in Agreement}}. \quad (1)$$

This equation is also appropriate for categorical variables that can take on more than two values, but we restrict our attention here to binary classifications. Maclure and Willett have recently discusssed additional issues that arise when classifications have three or more categories [7]. For binary classifications the expected proportion in agreement is $\pi_1\pi_2 + (1-\pi_1)(1-\pi_2)$, where π_1 is the proportion classified as having the characteristic according to the first classification and π_2 is the proportion classified as having the characteristic according to the second classification. Calculation of this expected proportion is analogous to the calculation of expected frequencies for the usual chi square test of association in a 2×2 table.

Kappa can be negative, but its value is 0.0 when agreement is at only the chance-expected level, and its value is 1.0 when agreement is perfect. As discussed below, a high value for kappa does not necessarily imply that one has accurately measured what was intended.

DEPENDENCE OF KAPPA ON PREVALENCE

Although kappa corrects for chance-expected agreement, this coefficient can be shown to depend not only on the sensitivity and specificity for each of the two classifications, but also on the true prevalence of the characteristic in the population studied. More specifically, if errors for the two binary classifications are assumed to be independent, then equation (1) may be expressed as follows:

$$\text{Kappa} = \frac{2\theta(1-\theta)(1-\alpha_1-\beta_1)(1-\alpha_2-\beta_2)}{\pi_1(1-\pi_2)+(1-\pi_1)\pi_2} \quad (2)$$

where θ is the true proportion having the characteristic, $1-\alpha_1 =$ specificity for the first classification, $1-\beta_1 =$ sensitivity for the first classification, $1-\alpha_2 =$ specificity for the second classification, and $1-\beta_2 =$ sensitivity for the second classification. The denominator of equation (2) can be expressed in terms of θ, α_1, β_1, α_2, and β_2 by noting that $\pi_1 = \theta(1-\beta_1) + (1-\theta)\alpha_1$ and $\pi_2 = \theta(1-\beta_2) + (1-\theta)\alpha_2$. Kraemer derived a similar expression, but only for the special case of $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$ [8]. Note that this formulation is a general one, in that it provides for perfect measurement ($\alpha = \beta = 1$), for worthless measurement ($\alpha = 1 - \beta$), and for everything in between.

In practice one does not generally know the true prevalence, sensitivity, or specificity. When these quantities are in fact available, the value of kappa is not of interest. Consequently, the importance of equation (2) stems not from any potential for application in actual studies but instead from the insight it provides into the dependence of kappa on the true prevalence of the characteristic. The strength of this dependence is illustrated in Fig. 1. For each of the three curves, we assume fixed sensitivities and specificities and illustrate how kappa would vary as a function of true prevalence. For curve A both measurements of the characteristic are assumed to have sensitivity of 95% and specificity of 99%; for curve B one measurement is assumed to have these same values for sensitivity and specificity but sensitivity of 70% and specificity of 90% for the other measurement; for curve C both measurements are assumed to have sensitivity of 70% and specificity of 90%. The figure illustrates that for fixed sensitivities

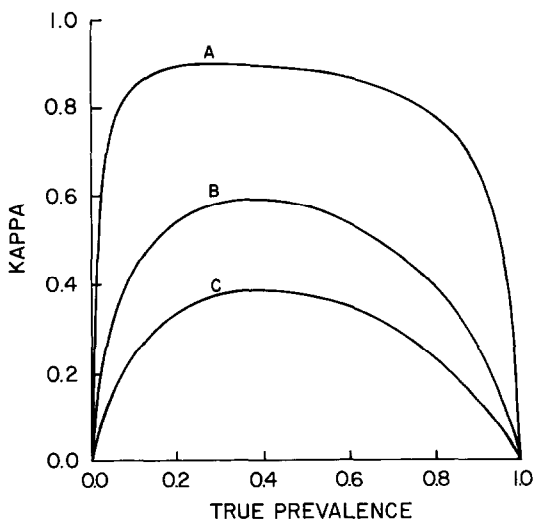

Fig. 1. Examples of the dependence of kappa on prevalence. For curve A both measurements of the characteristic are assumed to have sensitivity of 95% and specificity of 99%; for curve B one measurement is assumed to have these same values for sensitivity and specificity but sensitivity of 70% and specificity of 90% for the other measurement; for curve C both measurements are assumed to have sensitivity of 70% and specificity of 90%.

and specificities the value of kappa is highly dependent on the true prevalence. In fact, as the true prevalence approaches 0 or 1, the value of kappa necessarily approaches 0. An expression for the maximum value of kappa as a function of particular values of sensitivity and specificity is given in the Appendix.

This strong dependence of kappa on the true prevalence of the characteristic of interest complicates its interpretation as an index of the quality of measurement. It would seem especially difficult to compare two or more kappa values when the true prevalences for the groups or characteristics compared may differ. Thus, a population subgroup with the larger of two kappas may appear to reflect better measurement when in fact sensitivity and specificity could well be lower than for the other subgroup. Similarly, if ratings of one characteristic produce a higher kappa than do ratings for a second characteristic, one cannot be certain that the first characteristic is measured with any greater sensitivity and/or specificity than is the second.

As an illustration of this problem of interpretation, suppose that kappa is calculated for two symptoms, one of which is measured with 95% sensitivity and 99% specificity, whereas the other is measured with only 90% sensitivity and 98% specificity. Suppose also that the prevalence of the first symptom is 1%, whereas the prevalence of the second symptom is 40%. Figures 2 and 3 illustrate these two situations and indicate the values of kappa that would be obtained from studies of reliability for measurement of each of the symptoms. Independence of errors is again assumed. The upper 2×2 table in each figure is restricted to those who truly have the characteristic and the middle 2×2 table contains those who truly do not have the characteristic. The bottom 2×2 table is the composite of the other two, as would be observed in a reliability study, where the true status is unknown. For the symptom having a true prevalence of 1%, the value of kappa is shown in Fig. 2 to be 0.46. The corresponding value for the symptom having a true prevalence of 40% is 0.80. Consequently, the value of kappa for the second symptom is considerably greater than the value of kappa for the first symptom, even though the sensitivity and specificity for classification of the first symptom are both greater than for classification of the second symptom.

Symptom truly present

		Measurement 2 +	Measurement 2 −	
Measurement 1	+	0.009025	0.000475	0.0095
	−	0.000475	0.000025	0.0005
		0.0095	0.0005	0.0100

Symptom truly absent

		Measurement 2 +	Measurement 2 −	
Measurement 1	+	0.000099	0.009801	0.0099
	−	0.009801	0.970299	0.9801
		0.0099	0.9801	0.9900

Results from study of reliability

		Measurement 2 +	Measurement 2 −	
Measurement 1	+	0.009124	0.010276	0.0194
	−	0.010276	0.970324	0.9806
		0.0194	0.9806	1.0000

$$\text{Kappa} = \frac{0.009124 + 0.970324 - (0.0194)^2 - (0.9806)^2}{1 - (0.0194)^2 - (0.9806)^2} = 0.46$$

Fig. 2. Value of kappa from a reliability study for a symptom that has a true prevalence of 1% and that is measured with sensitivity of 95% and specificity of 99%.

Symptom truly present

Measurement 1	Measurement 2 +	Measurement 2 −	
+	0.324000	0.036000	0.3600
−	0.036000	0.004000	0.0400
	0.3600	0.0400	0.4000

Symptom truly absent

Measurement 1	Measurement 2 +	Measurement 2 −	
+	0.000240	0.011760	0.0120
−	0.011760	0.576240	0.5880
	0.0120	0.5880	0.6000

Results from study of reliability

Measurement 1	Measurement 2 +	Measurement 2 −	
+	0.324240	0.047760	0.3720
−	0.047760	0.580240	0.6280
	0.3720	0.6280	1.0000

$$\text{Kappa} = \frac{0.324240 + 0.580240 - (0.3720)^2 - (0.6280)^2}{1 - (0.3720)^2 - (0.6280)^2} = 0.80$$

Fig. 3. Value of kappa from a reliability study for a symptom that has a true prevalence of 40% and that is measured with sensitivity of 90% and specificity of 98%.

KAPPA AND THE ODDS RATIO

It has recently been suggested that the strong dependence of kappa on prevalence is a strength rather than a weakness [9]. Shrout *et al.* maintain that concern about this dependence, as expressed by Spitznagel and Helzer [4] in their attempt to identify a measure of agreement that is less influenced by prevalence, stems from confusion regarding the nature of reliability. We agree with Shrout *et al.* that the dependence of kappa on prevalence may in fact be a desirable property, but we reach this conclusion by a very different route. Our thinking is much more along the lines of that of Kraemer [8], who explored the relationship of kappa to statistical power and precision of estimation.

Although kappa is an index of agreement, or what is often called reproducibility or reliability, we feel that one of the important uses of such an index is to provide some indication—however qualified—of the likely impact of measurement error on the validity of indices of association such as the odds ratio. We propose that in etiologic research the quality and utility of a binary classification might be judged according to one particular criterion for validity, namely the degree to which non-differential misclassification [10, 11] attenuates the odds ratio for comparisons among groups. In this section we describe the relation between kappa and this criterion for validity.

Non-differential misclassification refers to those situations in which neither the sensitivity nor the specificity differs for two or more groups in a comparative study. Suppose, for example, that in a case–control study 40% of the cases have a particular exposure, as opposed to 20% of the controls. If the exposure could be classified without error, then the odds ratio would be 2.67. If, however, an imperfect measure having sensitivity of 80% and specificity of 95% is used, then the proportion of cases classified as exposed would be $(0.40)(0.80) + (1 - 0.40)(1 - 0.95) = 0.35$. The corresponding proportion for controls would be $(0.20)(0.80) + (1 - 0.20)(1 - 0.95) = 0.20$. The observed value of the odds ratio as calculated for the imperfectly measured exposure would then be $(0.35/0.65)/(0.20/0.80) = 2.15$. Because the misclassification is non-differential with respect to case–control status, it has the effect of attenuating the odds ratio from 2.67 to 2.15. The degree of this attenuation (i.e. bias toward

the null value) depends on the true prevalence of the characteristic as well as on sensitivity and specificity [10]. Therefore, it is of interest to know whether the kappa coefficient can, despite its dependence on prevalence, be interpreted as an index of the correspondence between the observed odds ratio and the odds ratio that would have been obtained had there been no misclassification. For this particular example, the anticipated value of kappa from a study of reliability for the measurement of exposure among cases would be 0.59 based on equation (2) and assuming $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$. The value of kappa among controls would be 0.56.

The measure we employ to quantify the quality of measurement when the true value of the odds ratio exceeds 1.0 is as follows:

$$\frac{\text{Observed Odds Ratio} - 1}{\text{True Odds Ratio} - 1} = \frac{\dfrac{\pi'(1-\pi)}{(1-\pi')\pi} - 1}{\dfrac{\theta'(1-\theta)}{(1-\theta')\theta} - 1} \quad (3)$$

where θ' and θ are the true prevalences within the study group and the reference group, respectively, and π' and π are the corresponding observed prevalences when the characteristic of interest is imperfectly measured. This ratio is interpretable as the proportion of the true effect that is captured when an imperfect technique of measurement is used. Its value is 0.0 when the association has been completely obliterated by essentially random classification, and its value is 1.0 when there is no misclassification and therefore no attentuation. For the case–control example, its value is $(2.15 - 1)/(2.67 - 1) = 0.69$.

Table 1 gives numerical results that indicate how well kappa reflects the attenuation of the odds ratio when misclassification is nondifferential with respect to membership in two groups (e.g. cases vs controls when a binary exposure is imperfectly measured). The value of kappa is calculated from equation (2) for the special case of $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$. We consider various combinations of values for sensitivity, specificity, and the true prevalence in the group we have designated as the reference group. The true value for the odds ratio is assumed to be 1.5 throughout and the true prevalence in the study group being compared to the reference group is set accordingly.

The results given in the table indicate that the value of kappa tends to be low when the measurement error would lead to substantial attenuation of the true odds ratio and that its value tends to be high when the observed odds ratio is close to the true odds ratio. For the particular 36 sets of parameter values considered, the correlation between kappa in the reference group and the measure (observed odds ratio $-$ 1)/(true odds ratio $-$ 1) is 0.98. The exact magnitude of this correlation would of course vary according to the specific parameter values examined. For these sets of values the mean kappa is slightly but significantly lower than the mean value of this measure of the correspondence between the observed and true odds ratios. Similar results were obtained when values other than 1.5 were considered for the true value of the odds ratio (not shown).

We have chosen to focus on the value of kappa in the reference group when assessing its usefulness as an index of the quality of measurement for estimation of the odds ratio. Our rationale is that reliability is generally assessed in a single group rather than in both of the groups that would ultimately be compared in an analytic study. Consider again a case–control study in which the two groups are compared in terms of an imperfectly measured binary exposure factor. The sort of data on reliability that might be available from some prior methodologic study would be kappa for a group with a true prevalence of exposure similar to that of the controls. Reliability studies among persons with the particular disease under study are less likely to be conducted. If in fact one did have available two imperfect measurements of a variable within each of the two groups being compared, then the true odds ratio would be directly estimable [12, 13] and kappa itself would be relatively superfluous.

Although kappa is most frequently used as a measure of reliability (i.e. agreement, reproducibility, consistency, or interchangeability of observers) in a single sample, the results in Table 1 indicate that under certain conditions kappa can also be interpreted in the context of validity of the odds ratio when comparing two samples. For fixed values of sensitivity and specificity, the variation of kappa with true prevalence tends to parallel differences in a straightforward index of the validity of a measurement technique for estimation of the odds ratio. Provided that misclassification is nondifferential, kappa seems to reflect quite well just how free from bias the odds ratio is likely to be, with a kappa of 0.0 indicating complete

Table 1. Kappa as an index of freedom from attenuation of an odds ratio of 1.5, according to true prevalence, sensitivity, and specificity

True prevalence in reference group	Sensitivity	Specificity	Kappa in study group	Kappa in reference group	Observed odds ratio − 1 / True odds ratio − 1
0.01	0.80	0.80	0.03	0.02	0.04
		0.95	0.14	0.10	0.14
		0.99	0.43	0.35	0.44
0.01	0.95	0.80	0.05	0.03	0.05
		0.95	0.20	0.14	0.16
		0.99	0.55	0.46	0.49
0.01	0.99	0.80	0.05	0.04	0.05
		0.95	0.22	0.16	0.17
		0.99	0.59	0.49	0.50
0.05	0.80	0.80	0.13	0.10	0.16
		0.95	0.41	0.33	0.44
		0.99	0.67	0.63	0.79
0.05	0.95	0.80	0.20	0.15	0.20
		0.95	0.54	0.45	0.50
		0.99	0.83	0.78	0.83
0.05	0.99	0.80	0.22	0.16	0.21
		0.95	0.57	0.48	0.51
		0.99	0.87	0.82	0.84
0.20	0.80	0.80	0.31	0.26	0.43
		0.95	0.59	0.56	0.73
		0.99	0.71	0.71	0.88
0.20	0.95	0.80	0.46	0.40	0.52
		0.95	0.77	0.73	0.81
		0.99	0.90	0.89	0.94
0.20	0.99	0.80	0.51	0.43	0.55
		0.95	0.82	0.78	0.83
		0.99	0.95	0.94	0.96
0.50	0.80	0.80	0.35	0.36	0.55
		0.95	0.54	0.58	0.71
		0.99	0.60	0.65	0.76
0.50	0.95	0.80	0.59	0.58	0.75
		0.95	0.80	0.81	0.88
		0.99	0.87	0.89	0.92
0.50	0.99	0.80	0.68	0.65	0.81
		0.95	0.89	0.89	0.94
		0.99	0.86	0.96	0.98

attenuation of the odds ratio and a value of 1.0 indicating a total lack of such bias.

When the two measurements from which kappa is calculated differ in quality ($\alpha_1 \neq \alpha_2$ or $\beta_1 \neq \beta_2$), the interpretation of kappa in terms of validity is more difficult. In general, for instance, use of the better of two measurements would result in an observed odds ratio that has been attenuated less than one would anticipate based on the value of kappa. Suppose, for example, that there are two imperfect methods available for classifying individuals in terms of a binary characteristic and that sensitivity and specificity are 0.90 and 0.99, respectively, for the first method vs 0.60 and 0.80 for the other method. The value of kappa in a group having a true prevalence of 0.10 is 0.22, as illustrated in Fig. 4. If the first method alone is employed for an analytic study in which the true prevalence in the reference group is 0.10 and in which the true odds ratio is 1.5, then the observed odds ratio would be 1.447 and the measure defined by equation (3) would be $(1.447 - 1)/(1.5 - 1) = 0.89$. Consequently, kappa may be a poor index of the quality of measurement when it is calculated from the agreement between measures of very unequal quality.

CORRELATED ERRORS

In all of the preceding discussion, we have assumed that the errors for the two imperfect measurements are uncorrelated, conditional on true status. Errors are uncorrelated if the sensitivity for the first measure is $1 - \beta_1$ regardless of whether or not particular true positives are correctly classified by the second measure and similarly if the specificity for the first measure is $1 - \alpha_1$ regardless of whether or not particular true negatives are correctly classified by the

Symptom truly present

Measurement 1	Measurement 2 +	Measurement 2 −	
+	0.054000	0.036000	0.0900
−	0.006000	0.004000	0.0100
	0.0600	0.0400	0.1000

Symptom truly absent

Measurement 1	Measurement 2 +	Measurement 2 −	
+	0.001800	0.007200	0.0090
−	0.178200	0.712800	0.8910
	0.1800	0.7200	0.9000

Results from study of reliability

Measurement 1	Measurement 2 +	Measurement 2 −	
+	0.055800	0.043200	0.0990
−	0.184200	0.716800	0.9010
	0.2400	0.7600	1.0000

$$\text{Kappa} = \frac{0.055800 + 0.716800 - (0.099)(0.24) - (0.901)(0.76)}{1 - (0.099)(0.24) - (0.901)(0.76)} = 0.22$$

Fig. 4. Value of kappa from a reliability study for a symptom that has a prevalence of 10% and that is measured by two methods, one with sensitivity of 90% and specificity of 99% and the other with sensitivity of 60% and specificity of 80%.

second measure. For example, in Fig. 2 the proportion of the population that truly does not have the symptom and is correctly classified as not having it by both of the two measurements was calculated as follows: $(1 - 0.01)(0.99)(0.99) = 0.970299$. When errors *are* correlated in that the first and second measurements tend systematically to misclassify the same individuals, the value of kappa is inflated so that it gives an overly reassuring impression of the quality of measurement. Again referring to Fig. 2, suppose that 0.5% rather than 0.0099% of the population consists of true negatives falsely classified as positive by both measurements. In that instance approximately equal numbers of true negatives would be falsely classified as positive by the first measurement only, by the second measurement only, and by both the first and second measurements. If the errors in classifying the true positives remained uncorrelated, this degree of correlation among the errors for the true negatives would result in a value of 0.72 for kappa, as calculated in Fig. 5, rather than the value of 0.46 as calculated in Fig. 2.

Correlated errors can occur for a number of reasons. One reason is that two imperfect measures may both tap the same extraneous characteristic. Another reason is that one measurement may not be made independently of the other. If the two measurements are classifications made by two experts, then a minimal requirement would be that each be kept blinded to the classifications of the other. Nevertheless, even if procedures consistent with maintaining independence are followed, multiple raters trained in a similar way may make similar errors. Also, the source data may themselves be flawed so that consistent application of rules for classification leads to high agreement despite the inadequacy of the data.

An important example of the problem of correlated errors involves the underlying cause of death as recorded on death certificates. Studies of agreement between multiple nosologists concerning the underlying cause of death have yielded impressive levels of agreement based on the data recorded on death certificates [14]. However, if the information recorded on the certificates is often inaccurate, then in terms of the actual cause of death the coded cause of

Sympton truly present

Measurement 1	Measurement 2 +	Measurement 2 −	
+	0.009025	0.000475	0.0095
−	0.000475	0.000025	0.0005
	0.0095	0.0005	0.0100

Symptom truly absent and with correlated errors

Measurement 1	Measurement 2 +	Measurement 2 −	
+	0.005000	0.004900	0.0099
−	0.004900	0.975200	0.9801
	0.0099	0.9801	0.9900

Results from study of reliability

Measurement 1	Measurement 2 +	Measurement 2 −	
+	0.014025	0.005375	0.0194
−	0.005375	0.975225	0.9806
	0.0194	0.9806	1.0000

$$\text{Kappa} = \frac{0.014025 + 0.975225 - (0.0194)^2 - (0.9806)^2}{1 - (0.0194)^2 - (0.9806)^2} = 0.72$$

Fig. 5. Value of kappa from a reliability study for a symptom that has a true prevalence of 1%, that is measured with sensitivity of 95% and specificity of 99%, and that is subject to correlated errors among the true negatives.

death may be grossly inadequate. Since multiple nosologists each code the same inaccurate certificates, consistent application of coding conventions leads to high levels of agreement because of correlated errors. Another example of correlated errors is in studies of test–retest reliability, where correlated errors often occur because a person who gives erroneous information on one occasion is more likely than other people to give erroneous information on a second occasion.

The problem of correlated errors raises a troublesome dilemma for the interpretation of the kappa coefficient. The best strategy for achieving independence of errors is to use two measurement techniques that are qualitatively different, e.g. reported symptoms vs radiographic evidence for a particular disease. Unfortunately, such efforts to ensure independence frequently lead to the comparison of measures having quite different sensitivities and specificities. As was discussed above, agreement involving two measurements of differing quality is particularly difficult to interpret.

The only practical advice we can offer in this regard is to avoid comparing measures that tap the same extraneous factors and to attempt to keep multiple imperfect measures as uncontaminated as possible from one another. If the result is two measures of widely different sensitivities and specificities, then it must be recognized that the kappa one calculates is likely to underestimate the value of (observed odds ratio − 1)/(true odds ratio − 1) for the better of the two measures. The likely magnitude of this underestimation is difficult to quantify unless one has at hand measures of validity that would render the calculations academic in the first place. As for evaluation of the poorer of the two measures, it may often be helpful to treat the better of the two measures as a criterion and to evaluate the poorer measure against it. For instance, if one followed this approach for the example in Fig. 4, estimates of 0.0558/0.099 = 0.564 and 0.1842/0.901 = 0.796 would be obtained for the sensitivity and specificity, respectively, of the second measure. Estimates such as these will be conservative in that they underestimate the true sensitivity and specificity, provided the errors for the two mea-

sures are uncorrelated. But, as in this example, the estimates are often not seriously biased.

CONCLUSION

Our reappraisal of the kappa coefficient in the context of potential attenuation of associations indicates that the strong dependence of kappa on true prevalence seems to be a desirable property if we elect to regard kappa as an index of validity of the odds ratio. The apparent inconsistency illustrated by the examples in Figs 2 and 3 proves not to be an inconsistency at all. Although sensitivity and specificity are both greater for the example in Fig. 2 than for the example in Fig. 3, it is entirely appropriate that the value of kappa for the second example should be substantially larger than the value for the first. This difference in kappas parallels the anticipated difference in the magnitudes of the observed vs true odds ratios for, say, comparing true prevalences of 2% vs the 1% in the first situation and 50% vs the 40% in the second situation.

Although we have considered the relationship between kappa and validity, it is important to note that the kappa we have studied is the kappa coefficient of *reliability*. That is, within the study group and within the reference group we have examined the agreement between two imperfect measurements of a variable subject to misclassification. This formulation contrasts with the case of "pure validity" considered by Spitznagel and Helzer [4], who examined a kappa-like measure calculated for true status vs a single imperfect measurement of that characteristic. We agree with Shrout *et al.* [9] that use of such a measure has little value because the true status is known.

The relationship of both kappa and the degree of attenuation of the odds ratio to true prevalence implies that prevalence is a critical consideration in the design of studies of reliability. In order for kappa to have a valid interpretation in terms of the magnitude that the observed vs true odds ratios would have in a comparative study, one must conduct a study of reliability within a population having a true prevalence of the imperfectly measured characteristic that is close in value to that for the population included within the comparative study. Frequently, it is tempting to conduct methodologic studies in samples of convenience or in groups known to have a greatly elevated prevalence of the characteristic of interest. Such practices may seem to enhance the efficiency of such studies, but they are clearly counterproductive when one desires a valid indicator of the quality of measurement for studies of associations in more representative groups.

Epidemiologists have demonstrated a healthy and increasing interest in the quality of the measurements they make. Recently, for instance, an extensive bibliography on observer variability was published in the **Journal of Chronic Diseases** [15]. During the 27 years since Cohen proposed the kappa coefficient, misgivings have surfaced only rather recently [2, 4]. We hope that our reappraisal of kappa in the context of the attenuation of association serves to clarify its interpretation and to counter what we regard as largely unwarranted skepticism about its utility.

REFERENCES

1. Carey G, Gottesman II. Reliability and validity in binary ratings: Areas of common misunderstanding in diagnosis and symptom ratings. **Arch Gen Psychiat** 1978; 35: 1454–1459.
2. Grove WM, Andreasen NC, McDonald-Scott P, *et al.* Reliability studies of psychiatric diagnosis: Theory and practice. **Arch Gen Psychiat** 1981; 38: 408–413.
3. Leckman JF, Sholomskas D, Thompson WD, *et al.* Best estimate of lifetime psychiatric diagnosis: A methodologic study. **Arch Gen Psychiat** 1982; 39: 879–883.
4. Spitznagel EL, Helzer JE. A proposed solution to the base rate problem in the kappa statistic. **Arch Gen Psychiat** 1985; 42: 725–728.
5. Fleiss JL. **Statistical Methods for Rates and Proportions.** New York: John Wiley; 1981: 2nd edn.
6. Cohen J. A coefficient of agreement for nominal scales. **Educ Psychol Meas** 1960; 20: 37–46.
7. Maclure M, Willett WC. Misinterpretation and misuse of the kappa coefficient. **Am J Epidemiol** 1987; 126: 161–169.
8. Kraemer HC. Ramifications of a population model for κ as a coefficient of reliability. **Psychometrika** 1979; 44: 461–472.
9. Shrout PE, Spitzer RL, Fleiss JL. Quantification of agreement in psychiatric diagnosis revisited. **Arch Gen Psychiat** 1987; 44: 172–177.
10. Copeland KT, Checkoway H, McMichael AJ, *et al.* Bias due to misclassification in the estimation of relative risk. **Am J Epidemiol** 1977; 105: 488–495.
11. Greenland S. The effect of misclassification in the presence of covariates. **Am J Epidemiol** 1980; 112: 564–569.
12. Thompson WD. Design issues in the assessment and control of misclassification errors. **Symposium presented at the Fifteenth Annual Meeting of the Society for Epidemiologic Research.** Cincinnati, Ohio, 16 June 1982.
13. Walter SD. Commentary on "Use of dual responses to increase validity of case-control studies". **J Chron Dis** 1984; 37: 137–139.
14. Curb JD, Babcock C, Pressel S, *et al.* Nosological coding of cause of death. **Am J Epidemiol** 1983; 118: 122–128.

15. Feinstein AR. A bibliography of publications on observer variability. **J Chron Dis** 1985; 38: 619–632.

APPENDIX

The maximum value of kappa for given values of α_1, β_1, α_2, and β_2 is attained at a unique value of the true prevalence (θ). In order to obtain an expression for this value of θ, one can take the logarithm of the expression in equation (2) and differentiate with respect to θ, yielding:

$$\theta = \frac{[\alpha_1(1-\alpha_2)+(1-\alpha_1)\alpha_2]^{1/2}}{[\alpha_1(1-\alpha_2)+(1-\alpha_1)\alpha_2]^{1/2} + [\beta_1(1-\beta_2)+(1-\beta_1)\beta_2]^{1/2}}. \quad \text{(A.1)}$$

Substitution of this value for θ into equation (2) leads to the following expression for the maximum value of kappa:

$$\text{kappa}_{\max} = \frac{2(1-\alpha_1-\beta_1)(1-\alpha_2-\beta_2)}{2-\alpha_1(1-\beta_2)-(1-\alpha_1)\beta_2-\alpha_2(1-\beta_1)-(1-\alpha_2)\beta_1 + 2\{[\alpha_1(1-\alpha_2)+(1-\alpha_1)\alpha_2] \times [\beta_1(1-\beta_2)+(1-\beta_1)\beta_2]\}^{1/2}}. \quad \text{(A.2)}$$

Kraemer derived similar equations for the special case of $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$ [7]. For the examples shown in Fig. 1, the maximum value of kappa is attained for curves A, B, and C when the true prevalence is 0.313, 0.367, and 0.396, respectively. The maximum value of kappa in these three instances is 0.898, 0.585, and 0.385.