J Clin Epidemiol Vol. 43, No. 6, pp. 551–558, 1990
Printed in Great Britain.

0895-4356/90 $3.00 + 0.00

HIGH AGREEMENT BUT LOW KAPPA: II. RESOLVING THE PARADOXES*

DOMENIC V. CICCHETTI† and ALVAN R. FEINSTEIN‡

Yale University School of Medicine, 333 Cedar Street, P.O. Box 3333, New Haven, CT 06510, U.S.A.

(*Received in revised form* 20 *November* 1989)

Abstract—An omnibus index offers a single summary expression for a fourfold table of binary concordance among two observers. Among the available other omnibus indexes, none offers a satisfactory solution for the paradoxes that occur with p_0 and κ. The problem can be avoided only by using p_{pos} and p_{neg} as two separate indexes of proportionate agreement in the observers' positive and negative decisions. These two indexes, which are analogous to sensitivity and specificity for concordance in a diagnostic marker test, create the paradoxes formed when the chance correction in κ is calculated as a product of the increment in the two indexes and the increment in marginal totals. If only a single omnibus index is used to compared different performances in observer variability, the paradoxes of κ are desirable since they appropriately "penalize" inequalities in p_{pos} and p_{neg}. For better understanding of results and for planning improvements in the observers' performance, however, the omnibus value of κ should always be accompanied by separate individual values of p_{pos} and p_{neg}.

Kappa Concordance Agreement Paradox

INTRODUCTION

In Part I [1], we noted how imbalances in the distribution of marginal totals can sometimes produce two types of paradoxes when the variability of two observers for binary decisions is expressed with the κ coefficient and with p_0, the proportion of observed agreement. κ can sometimes be low despite relatively high values of p_0; and κ values will sometimes be increased, rather than decreased, by departures from symmetry in the vertical and horizontal marginal totals of the 2×2 concordance table. In this paper, we determine whether the problems can be eliminated with omnibus indexes other than κ, and we propose a method of resolving the problems.

OTHER OMNIBUS INDEXES

An omnibus index, such as p_0 or κ, offers a single expression that summarizes the results of a 2×2 table of concordance. In studies of diagnostic marker tests, rather than observer variability, two omnibus expressions have been used in the past. One of these, the "index of validity", corresponds exactly to p_0, the proportion of observed agreement. The other omnibus index, Youden's J [2], is calculated from two individual indexes, *sensitivity* and *specificity*. Sensitivity is the proportionate accuracy of the

*Supported in part by a Grant from the Andrew W. Mellon Foundation and by Veterans Administration Funds.
†Senior Research Psychologist and Biostatistician, Veterans Administration Medical Center, West Haven, Connecticut; and Senior Research Scientist, Department of Psychiatry and Child Study Center, Yale University School of Medicine, New Haven, Connecticut.
‡Professor of Medicine and Epidemiology; Director, Clinical Epidemiology Unit and The Robert Wood Johnson Clinical Scholars Program, Yale University School of Medicine, New Haven, Connecticut. Senior Biostatistician, Cooperative Studies Program Coordinating Center, Veterans Administration Medical Center, West Haven, Connecticut.

diagnostic test in identifying positive cases of disease; and specificity is the corresponding proportionate accuracy in identifying the negative control group. Youden's J has an algebraic structure that makes it equal the sum of sensitivity plus specificity minus 1. The values of κ and J will be identical, as shown by Kramer [3], when the two observers have equal marginal totals in their ratings, i.e. $f_1 = g_1$ and $f_2 = g_2$.

Beyond p_0 and κ, the omnibus indexes that have been proposed for tests of concordance (rather than accuracy) can be divided into those that do or do not contain overt corrections for chance agreement.

1. *Non-chance-corrected indexes*

The omnibus indexes that do not overtly correct for chance are Yule's Y, Maxwell's RE, and the odds ratio.

Yule's Y. In a fourfold table of the form

$$\begin{Bmatrix} a & b \\ c & d \end{Bmatrix},$$

Yule originally proposed [4] an index of association calculated as

$$Y = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}.$$

This index does not contain an overt chance correction, although its value should be 0 for completely chance agreement and 1 for perfect agreement. The index has received relatively little usage in recent years, because other indexes (such as the odds ratio or the ϕ correlation coefficient) have been much more popular as omnibus indexes of association. In 1985, however, Spitznagle and Helzer [5] advocated the use of Yule's Y as a concordance index that could overcome the "base rate problem" (i.e. unbalanced marginal totals) of the κ statistic. Shrout *et al.* [6] later pointed out, however, that the "base rate" problem is not avoided if f_1 and g_1 are very large or small relative to f_2 and g_2. In addition, Y has two overwhelming disadvantages: it measures association, not agreement; and the square root transformations have no recognizable intuitive meaning.

Maxwell's RE. Maxwell's Random Error (RE) statistic was intended [7] to eliminate the paradox of low κ with high agreement. The basic idea was to create a κ-like index that had an "invisible" chance agreement level of 0.5. Thus, if p_0 is 0.8, the value of RE would be $(0.8 - 0.5)/(1 - 0.5) = 0.60$. With this formula,

$$RE = \frac{p_0 - 0.5}{1 - 0.5} = 2p_0 - 1.$$

The same result can be obtained by introducing $p_d = (b + c)/N$ as an index of proportional disagreement, and then defining

$$RE = p_0 - p_d.$$

Since $RE = 2p_0 - 1$, it contains no apparent advantages over p_0 alone, and offers no particular mathematical repairs for the cited paradoxes.

Odds ratio. In a 2×2 table of the form

$$\begin{Bmatrix} a & b \\ c & d \end{Bmatrix},$$

the odds ratio is ad/bc. If this is a concordance table, a and d represent the cells of agreement, and b and c represent disagreements. Although not overtly corrected for chance, increases in the value of ad/bc will represent increasingly high agreement, and purely chance agreement should yield a value of 1.

In epidemiologic research, the odds ratio has been particularly valuable, because—in suitable circumstances—it can be used as a surrogate for the risk ratio. Despite this advantage in epidemiologic studies of association, however, the odds ratio is not an attractive index of concordance. The odds ratio has no intuitive meaning for agreements; it cannot be immediately calculated if one of the b or c cells is zero; and (despite possibly high values for p_0 or κ) the odds ratio will give a falsely low result of zero if either the a or d cell is zero.

2. *Chance-corrected indexes*

Regardless of whether Yule's Y, Maxwell's RE, or the odds ratio is inherently satisfactory as an omnibus index of concordance, none of these three indexes contains an overt correction for chance agreement. The omnibus indexes that contain such corrections are the ϕ coefficient, and several other indexes that are essentially equivalent to κ.

ϕ Coefficient. The ϕ coefficient of correlation in a 2×2 table is obtained as $\sqrt{X^2/N}$ where χ^2 is calculated in the traditional chi-square manner as $(ad - bc)^2 N/f_1 f_2 g_1 g_2$. The value of ϕ will be $\phi = (ad - bc)/\sqrt{f_1 f_2 g_1 g_2}$. The correction for chance is produced because X^2 is constructed as a sum of the (observed − expected)^2/expected values in each of the four cells of the table.

The structure of ϕ has a reasonable resemblance to the structure of κ. With some algebraic transformations that will not be displayed here, it can be shown that

$$\kappa = \frac{ad - bc}{\left(\frac{f_1 g_2 + f_2 g_1}{2}\right)}.$$

Since ϕ has the same numerator as κ, but has $\sqrt{f_1 f_2 g_1 g_2}$ in its denominator, the κ and ϕ coefficients will usually produce reasonably similar results. This point can be demonstrated using the symbols of our previous paper [1], in which $f_1 = vN$, $f_2 = (1-v)N$, $g_1 = wN$, $g_2 = (1-w)N$. With these symbols,

$$\frac{f_1 g_2 + f_2 g_1}{2} = \frac{v(1-w)N^2 + w(1-v)N^2}{2}$$

$$= \frac{N^2}{2}[v + w - 2vw]$$

and

$$\sqrt{f_1 f_2 g_1 g_2} = N^2\sqrt{vw(1-v)(1-w)}.$$

The denominator of κ^2 will be $N^4(v + w - 2w)^2/4$ and the denominator of ϕ^2 will be $N^4[vw(1-v)(1-w)]$. Some algebra not shown here will demonstrate that the denominator for κ^2 exceeds the denominator of ϕ^2 by the quantity $N^4(v-w)^2/4$. Extending the algebra, we can express

$$\phi^2 = \kappa^2\left[1 + \frac{N^4(v-w)^2}{4f_1 f_2 g_1 g_2}\right]$$

ϕ^2 will equal k^2 when $v = w$, a situation in which the marginal totals are either symmetrical, with $f_1 = g_1$ and $f_2 = g_2$, or perfectly balanced, with $f_1 = f_2 = g_1 = g_2$. In other circumstances, since v and w are each < 1, $v - w$ will always be < 1. The small value of $v - w$ will be even smaller when squared. Thus, $N^4(v-w)^2$ will usually be much less than $4f_1 f_2 g_1 g_2$, and their ratio will usually be $\ll 1$. ϕ^2 will therefore usually be only slightly larger than κ^2.

For example, in Table 1 of Part I of this paper [1], $N^4(v-w)^2 = 100^4(0.46 - 0.49)^2 = 90{,}000$ and $4f_1 f_2 g_1 g_2 = 24{,}830{,}064$. Their ratio is 0.0036, so that the difference of ϕ^2 and κ^2 is tiny. (In the actual calculations, $\kappa = 0.6995$ and $\phi^2 = 0.7008$ for that table.)

For the slight asymmetry of the unbalanced marginal totals of Part I, Table 2, $\phi = 0.33$ and $\kappa = 0.32$. In Part I, Table 3, $\phi = 0.13$, and $\kappa = 0.13$. In Part I, Table 4, where the asymmetrical unbalanced marginal totals produced a paradoxical rise in κ, the effect of ϕ is more distinctive, although still relatively small. For that table, $\phi = 0.312$ and $\kappa = 0.259$. In Table 5, $\phi = \kappa = 0.44$, as would be expected from the fact that $f_1 = g_1$ and $f_2 = g_2$. In Table 6, $\phi = 0.76$ and $\kappa = 0.74$.

In most situations, therefore, ϕ will not be substantially different from κ. In particular, ϕ will not resolve the κ paradoxes. In the first type of paradox, v and w will both be high (or low), and the values for κ and ϕ will be quite close. In the second paradox, when v and w lie on opposite sites of 0.5, ϕ will increase the paradox by further elevating the value of κ.

Consequently, ϕ does not offer solutions to the κ paradoxes. Besides, as an index of association between two *different* variables, ϕ does not have good intuitive meaning as an index of concordance for two measurements of the *same* variable.

Other expressions. Several other omnibus indexes have also been developed for expressing chance-corrected observer agreement in a fourfold table. They are the "standard deviation agreement" index of Armitage *et al.* [8], Rogot and Goldberg's "agreement index" [9], and Goodman and Kruskal's lambda [10]. None of these indexes is particularly distinctive, however. Fleiss [11] has shown that, with only minor modifications, they are all mathematically equivalent to κ.

THE RATIONALE FOR AN OMNIBUS INDEX

In view of the difficulties of getting a satisfactory replacement for the paradoxes of κ, another level of inquiry is to reappraise the desirability of a single omnibus index. Why do we *want* a single index?

If diagnostic marker tests require two indexes of evaluation (sensitivity and specificity), why should a single index be used for analogous studies of observer variability? Probably the main or only value of the omnibus index is that it offers a single statistical summary for a study of concordance. This single summary can suffice if it is the only goal in the research. If we really want to *use* the results of the research, however, *no* omnibus index can be satisfactory.

This problem has already been noted in diagnostic marker tests, where omnibus indexes are almost never cited today because they cannot be used effectively. Separate indexes of sensitivity and specificity have become popular and have replaced omnibus indexes because the user of a

diagnostic marker test wants to know what is connoted when the test is positive and when the test is negative. A single omnibus index does not answer either question.

Analogously, someone who wants to use the results of a study of observer variability will want to know what was found for agreement in positive and negative responses. The user might even want to know further what is connoted when the positive or negative response is given by Observer A or Observer B. No matter how a single omnibus index is constructed, the distinctions in positive or negative interpretations will be lost when everything is "lumped" together for the single omnibus value.

SEPARATE INDEXES OF AGREEMENT

A simple way to resolve the omnibus problem is to cite separate indexes, analogous to sensitivity and specificity, for the positive and negative agreements noted in an observer variability study. Since neither observer is regarded as the "gold standard", proportions of agreement can be calculated for the average of their positive and their negative responses.

Positive agreement

Three indexes can be used to express positive agreement. They are called p_{pos}, corrected p_{pos}, and Chamberlain's p_{pos}.

p_{pos}. The observed proportion of positive agreement can be designated as p_{pos}. If the number of positive readings is f_1 for Observer A and g_1 for Observer B, the average number of positive readings is $(f_1+g_1)/2$. The index of average positive agreement would be

$$p_{pos}=\frac{a}{\left(\frac{f_1+g_1}{2}\right)}=\frac{2a}{f_1+g_1}.$$

This index, unlike p_0, depends exclusively on uni-directional decisions. Since the denominator value of $(f_1+g_1)/2$ shows how many of these decisions were made, a correction for chance agreement seems less necessary than for p_0, which obscures differences in the two sets of denominators.

Corrected p_{pos}. If desired, a chance correction for p_{pos} can be calculated using the same ratio strategy as κ. The numerator of the ratio is the proportion of observed agreement minus the expected proportion of chance agreement. The denominator is formed by the proportions of perfect agreement minus chance agreement. Since the expected number of agreements in the a cell is f_1g_1/N, the proportion of expected agreement is $[f_1g_1/N]/[f_1+g_1)/2]$. The ratio for corrected p_{pos} would be

$$\left[\frac{2a}{f_1+g_1}-\frac{2f_1g_1}{N(f_1+g_1)}\right]\Big/\left[1-\frac{2f_1g_1}{N(f_1+g_1)}\right].$$

After the algebra is worked out, this becomes

$$\text{corrected } p_{pos}=\frac{2aN-2f_1g_1}{N(f_1+g_1)-2f_1g_1}.$$

As Fleiss [11] has pointed out, however, this value of corrected p_{pos} in a 2×2 table is mathematically identical to κ itself. Therefore, the calculation of corrected p_{pos} offers no additional information.

Chamberlain's p_{pos}. Chamberlain *et al.* [12] proposed an index of proportionate positive agreement in which the actual positive agreement is divided by *all* positive readings made by either observer. This index would be

$$p_{ppa}=\frac{a}{a+b+c}=\frac{a}{N-d}.$$

The proponents believed that this expression would be particularly useful in binary assessments of observer variability when the positive event has a low prevalence, i.e. for relatively large values of d.

The relationship between p_{pos} and p_{ppa} can be noted by substituting $f_1=a+c$, $g_1=a+b$, and $f_1+g_1=2a+b+c=N-d+a$ into the formula for p_{pos}, so that

$$p_{pos}=\frac{2a}{N+(a-d)}.$$

With some algebra not shown here, it turns out that

$$p_{ppa}=\frac{p_{pos}}{2-p_{pos}}.$$

Because of the direct conversion between p_{pos} and p_{ppa}, the Chamberlain p_{ppa} does not appear particularly useful.

Negative agreements

In direct correspondence to the foregoing approach for expressing positive agreement, indexes of average proportional negative agreement can be calculated as

$$p_{neg}=\frac{2d}{f_2+g_2}=\frac{2d}{N-(a-d)},$$

and

$$\text{corrected } p_{neg} = \frac{2dN - 2f_2g_2}{N(f_2+g_2) - 2f_2g_2}.$$

The negative counterpart of Chamberlain's coefficient would be

$$p_{pna} = \frac{d}{b+c+d} = \frac{d}{N-a}.$$

and

$$d = \frac{Np_{neg}(1-p_{pos})}{2-p_{pos}-p_{neg}}.$$

These values can then be substituted for a and d in formula (A.1) of Part I [1]. When the algebra is worked out, the value of κ becomes

$$\kappa = \frac{N^2(p_{pos}+p_{neg}-2p_{pos}p_{neg}) - 2f_1f_2(2-p_{pos}-p_{neg}) - N(f_1-f_2)(p_{pos}-p_{neg})}{N^2(2-p_{pos}-p_{neg}) - 2f_1f_2(2-p_{pos}-p_{neg}) - N(f_1-f_2)(p_{pos}-p_{neg})}. \quad (6)$$

For the same reasons discussed earlier for indexes of positive agreement, p_{neg} is the only substantially valuable member of the indexes of negative agreement. The corrected p_{neg} is identical to κ in a 2×2 table; and p_{pna} is an algebraic transformation of p_{neg}.

ROLE OF p_{pos} AND p_{neg} IN p_0 AND KAPPA

As indexes of the "basic" agreement in each direction of decision, p_{pos} and p_{neg} have fundamental roles in describing the performance of the two observers, and also in the statistical calculations of both p_0 and κ.

Role in p_0

The value of p_0 is a weighted sum of the values for p_{pos} and p_{neg}. The "weights" are obtained from the marginal proportions of positive and negative readings. Thus, for the total of $2N$ readings (consisting of N readings from each observer), the proportion of positive readings is $(f_1+g_1)/2N$. The proportion of negative readings is $(f_2+g_2)/2N$. With these "weights",

$$p_0 = (p_{pos})\left(\frac{f_1+g_1}{2N}\right) + (p_{neg})\left(\frac{f_2+g_2}{2N}\right),$$

which becomes $(a+d)/N$, when the algebra is carried out.

Role in kappa

The role of p_{pos} and p_{neg} in κ requires some further algebraic development. The formulas $p_{pos} = 2a/[N+(a-d)]$ and $p_{neg} = 2d/[N-(a-d)]$ can become a pair of simultaneous equations, which can be solved for a and d. The results will show that

$$a = \frac{Np_{pos}(1-p_{neg})}{2-p_{pos}-p_{neg}}$$

In the numerator and denominator of this expression, the terms $N^2(p_{pos}+p_{neg}-2p_{pos}p_{neg})$ and $N^2(2-p_{pos}-p_{neg})$ will always be positive, and the term that subtracts $2f_1f_2(2-p_{pos}-p_{neg})$ will always be negative. The term in $(f_1-f_2)(p_{pos}-p_{neg})$ will be a "swing" term, analogous to $(f_1-f_2)(a-d)$ in formula (A.1) of Part I [1]. Because f_1 contains a and f_2 contains d, a concordant increase (or decrease) in both f_1 and a should ordinarily yield a positive product for $(f_1-f_2)(p_{pos}-p_{neg})$. Since this product is subtracted, however, κ can be raised if the product is negative and lowered if positive. The sign and magnitude of the change produced by this product can lead to paradoxical decreases or increases of κ in the previously illustrated examples. For the six tables in Part I [1], the corresponding results are tabulated below:

Previous Table	p_0	κ	p_{pos}	p_{neg}	$p_{pos}-p_{neg}$	f_1-f_2
1	0.85	0.70	0.842	0.857	−0.015	−8
2	0.85	0.32	0.914	0.400	0.514	70
3	0.60	0.13	0.692	0.429	0.263	40
4	0.60	0.26	0.555	0.636	−0.081	−40
5	0.90	0.44	0.944	0.500	0.444	80
6	0.90	0.74	0.933	0.800	0.133	40

In each of the six tables, f_1-f_2 and $p_{pos}-p_{neg}$ have similar signs, so that $(f_1-f_2)(p_{pos}-p_{neg})$ will be positive. In both Tables 1 and 2, p_0 was 0.85. In Table 1, however, p_{pos} and p_{neg} were almost equal (0.842 and 0.857); whereas in Table 2, $p_{pos}-p_{neg}$ was 0.514. Since $f_1 > f_2$ in Table 2, the net effect was the dramatic lowering of κ that produced its paradox for that table. In both Tables 3 and 4, p_0 was 0.60. The difference of 0.263 between p_{pos} and p_{neg}, however, was larger in Table 3 than the −0.081 difference in Table 4. Since $f_1 > f_2$ in Table 3 but $f_1 < f_2$ in Table 4, the net effect of the products was to reduce κ in both tables, but the reduction was greater in Table 3.

In Table 5, $p_{pos} - p_{neg}$ showed a large difference of 0.444 and $f_1 - f_2$ was also relatively high at 80. Their product reduced κ much more than the corresponding product, in Table 6, where $p_{pos} - p_{neg} = 0.133$ and $f_1 - f_2 = 40$. Another important point about Tables 2 and 6 is their unacceptably low values of p_{neg} (at 0.400 and 0.500 respectively) despite the high values of 0.85 and 0.90 for p_0.

The striking effect of the product of $(f_1 - f_2)(p_{pos} - p_{neg})$ indicates the importance of noting each of the main two sources of the κ paradoxes. The value of $f_1 - f_2$ reflects the marginal total imbalances that were emphasized in our previous paper. The value of $p_{pos} - p_{neg}$, as emphasized now, reflects the direct performance of the two observers. The structure of κ in formula (6) indicates that the best way to avoid erratic or paradoxical behavior in κ is for the term $(f_1 - f_2)(p_{pos} - p_{neg})$ to be zero. One way to achieve this goal is if $f_1 = f_2$, when the marginal totals have a balanced distribution. This achievement is usually attainable only if the investigator can design the study by choosing approximately equal numbers of positive and negative challenge cases. If the status of those cases is not known in advance, however, or if the investigator wants the challenge to reflect the clinical realities of the actual distribution of available cases, f_1 and f_2 may be substantially unequal.

The other way to avoid the κ paradoxes, therefore, is for the observers to do their work in a way that makes $p_{pos} - p_{neg}$, so that the observers' agreement is "balanced" for positive and negative components. If $p_{pos} = p_{neg} = p'$, the foregoing formula (6) for κ reduces to

$$\kappa = \frac{p'N^2 - 2f_1f_2}{N^2 - f_1f_2}.$$

Also, if $p_{pos} = p_{neg} = p'$, the values of a and d will be equal. Because $p_{pos} = 2a/(2a + b + c)$ and $p_{neg} = 2d/(2d + b + c)$, both formulas share the same values of b and c, and equalities in p_{pos} and p_{neg} cannot be achieved unless $a = d$. Consequently, whenever $a \neq d$ in a fourfold concordance table, inequality can be expected for p_{pos} and p_{neg}, unless the interobserver agreement is perfect. When $a = d$, the previous formula (A.1) for κ will become

$$\kappa = \frac{2aN - 2f_1f_2}{N^2 - f_1f_2}.$$

The value of $p_{pos} = p_{neg} = p'$ will be the same as $p_0 = 2a/N$.

Importance of p_{pos} and p_{neg}

The individual values of p_{pos} and p_{neg} make three important contributions when results are interpreted for study of observer variability. First, p_{pos} and p_{neg} will indicate the consistency of the two observers when going in the opposite directions of positive and negative decisions. The distinction will help a reader decide about the persuasiveness of the individual results, and will also help the investigator design further work, if needed, to decrease the observers' disparities in the positive direction, negative direction, or both.

A second contribution of the p_{pos} and p_{neg} values is that they can explain and "eliminate" the paradox of high p_0 but low κ. The paradox arises because of the effect of symmetrically unbalanced marginal totals when p_0 is converted to κ. If the main weight of the imbalance is in a positive direction and if p_{pos} is high, the effect of a low p_{neg} will be obscured in p_0. (The same effect will occur with a low p_{pos}, a high p_{neg}, and a negatively weighted imbalance.) The first κ paradox occurs because of the "penalty" produced by the correction for chance-expected agreement. If the disparate p_{pos} and p_{neg} values are directly displayed, however, the discrepancy is immediately evident, and its existence can be recognized without the need for a signal from the unexpectedly lowered κ.

A third contribution of the p_{pos} and p_{neg} values occurs in the second κ paradox, which is produced by less than perfect symmetry or by striking asymmetry in balance of the marginal totals. In this situation, κ "rewards" the observers who manage, despite the asymmetry, to get the same p_0 as in more symmetrical balances. If p_0 is high, but either p_{pos} or p_{neg} is low, the κ correction makes an appropriate downward adjustment for the poorer performance. If p_0 is not high enough to be attractive, κ will seldom be elevated into a more attractive range. When neither p_0 nor κ is particularly high, the investigator will seldom want to claim that the observers have produced laudable results for agreement. Instead, the investigator's next step will usually be to try to improve the performance of the observational system. The values of p_{pos} and p_{neg} will be helpful guides to indicating where and how much improvement is needed.

CONCLUSIONS AND RECOMMENDATIONS

In correcting the value of p_0 for chance agreement, the calculation of κ contains a crucial component that is a product of two increments. One of these increments is $p_{pos} - p_{neg}$ (or $a - d$). It represents the disparity in the observers' agreements for negative and positive ratings. The other increment, $f_1 - f_2$, represents the imbalance in marginal totals of the 2×2 concordance table.

With changes in the relative magnitudes of these two sets of increments, their product can substantially raise or lower the value obtained when the observer's proportion of total agreement is converted to κ. The results can lead to two types of paradox. In one paradox, a large, reasonably symmetrical imbalance in the marginal totals can convert a reasonably high p_0 into a much smaller value of κ. In the other paradox, κ values for the same p_0 can be unexpectedly raised when imbalances in the marginal totals are *not* perfectly symmetrical.

Because the marginal total values of $f_1 - f_2$ are essentially determined in advance by the investigator's choice of the population under study, the main source of the paradoxes will be observers' separate performances for p_{pos} and p_{neg}. These separate distinctions in performance will be lost in any single omnibus index, such as p_0 or κ. Nevertheless, the κ index, although seemingly "unfair" to the observers when a paradox is created, actually has a desirable scientific function. κ will penalize gross disparities in p_{pos} and p_{neg} that may be deceptively obscured in p_0 alone if the marginal totals are not symmetrically balanced; and κ will reward fair (e.g. 70–79%) to good (e.g. 80–89%) values of p_0 that are obtained despite the handicap of asymmetrical imbalances in the marginal totals.

In both circumstances, the correction factor in κ, although originally intended to deal with agreements that might occur by chance, does a valuable scientific job in adjusting the results for discrepancies in the separate distinctions of p_{pos} vs p_{neg} and in the marginal totals of f_1 vs f_2.

An ideal statistical summary of results for a concordance study will therefore depend on the purpose of the summary. If the research has been completed, and if the results are to be compared with those of analogous studies of observer variability, κ is the preferred single index of expression. Its apparent paradoxes are the inevitable and desirable result of a correction factor that helps adjust or "standardize" disparities in the populations under study and in performance of the observers. In this way, κ is analogous to the single-index standardizations [13] that adjust analogous disparities in other forms of epidemiologic data, and that also have the potential for producing paradoxical results. For the edification of readers, however, the publication of κ should always be accompanied by individual values of p_{pos} and p_{neg}.

On the other hand, in most research circumstances the purpose of studying observer variability is to reduce it, rather than quantify its magnitude. If p_0 is high (e.g. $\geq 90\%$), the first and second κ paradoxes—as illustrated in our previous Tables 5 and 6—can suggest much worse agreement than what actually occurred. If p_0 is not particularly high, the second paradox may elevate κ but will seldom raise it to an impressively good level. In either situation, the investigator's main goal will usually be improve the sub-optimal performance of the observers, rather than to publish results that "expose" their relatively poor agreement.

For purposes of understanding what happened in the observational process, or improving the observers' agreement, no single omnibus index will be satisfactory. The results of p_{pos} and p_{neg} must be examined separately, to avoid any obscuring or deceptive effects that may be produced by p_0, κ, or any other omnibus index. An analogous discontent with omnibus indexes has led to the use of two separate indexes, sensitivity and specificity, in reporting the results of concordance studies for diagnostic marker tests. Unlike the latter two indexes, however, p_{pos} and p_{neg} do not have a relatively reciprocal relationship in which one value must rise as the other falls. The values of p_{pos} and p_{neg} can each range freely between 0 and 1, according to the performance of the observers.

Our conclusions and recommendations, therefore, are that the paradoxes of κ arise from a desirable adjustment. It helps "standardize" and enhance the value of κ as a single omnibus index for comparing results in different studies of observer variability. For better understanding of individual results, however, and for planning improvements in the observers' performance, the omnibus value of κ should always be accompanied by separate individual values of p_{pos} and p_{neg}.

REFERENCES

1. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. **J Clin Epidemiol** 1990; 43: 543–549.

2. Youden WJ. Index for rating diagnostic tests. **Cancer** 1950; 3: 32–35.
3. Kraemer HC. Estimating false alarms and missed events from interobserver agreement: Comment on Kaye. **Psychol Bull** 1982; 92: 749–754.
4. Yule GU. On the methods of measuring association between two attributes. **J R Stat Soc** 1912; 75: 579–642.
5. Spitznagel EL, Helzer JE. A proposed solution to the base rate problem in the kappa statistic. **Arch Gen Psychiat** 1985; 42: 725–728.
6. Shrout PE, Spitzer RL, Fleiss JL. Quantification of agreement in psychiatric diagnosis revisited. **Arch Gen Psychiat** 1987; 44: 172–177.
7. Maxwell AE. Coefficients of agreement between observers and their interpretation. **Br J Psychiat** 1977; 130: 79–83.
8. Armitage P, Blendis LM, Smyllie HC. The measurement of observer disagreement in the recording of signs. **J R Stat Soc Ser A** 1966; 129: 98–109.
9. Rogot E, Goldberg I. A proposed index for measuring agreement in test–retest studies. **J Chron Dis** 1966; 19: 991–1006.
10. Goodman LA, Kruskal WH. Measures of association for cross classifications. **J Am Stat Soc** 1954; 49: 732–764.
11. Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait. **Biometrics** 1975; 31: 651–659.
12. Chamberlain J, Ginks S, Rogers P, Nathan BE, Price JL, Burn I. Validity of clinical examination in mammography as screening tests for breast cancer. **Lancet** 1975; 2: 1026–1030.
13. Chan CK, Feinstein AR, Jekel JK, Wells CK. The value and hazards of standardization in clinical epidemiologic research. **J Clin Epidemiol** 1988; 41: 1125–1134.