STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:2109–2129 (DOI: 10.1002/sim.1180)

TUTORIAL IN BIOSTATISTICS
Kappa coefficients in medical research

Helena Chmura Kraemer[1,*,†], Vyjeyanthi S. Periyakoil[2] and Art Noda[1]

[1]*Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, California, U.S.A.*
[2]*VA Palo Alto Health Care System, Palo Alto, CA, U.S.A.*

SUMMARY

Kappa coefficients are measures of correlation between categorical variables often used as reliability or validity coefficients. We recapitulate development and definitions of the K (categories) by M (ratings) kappas ($K \times M$), discuss what they are well- or ill-designed to do, and summarize where kappas now stand with regard to their application in medical research. The $2 \times M$ ($M \geqslant 2$) intraclass kappa seems the ideal measure of binary reliability; a 2×2 weighted kappa is an excellent choice, though not a unique one, as a validity measure. For both the intraclass and weighted kappas, we address continuing problems with kappas. There are serious problems with using the $K \times M$ intraclass ($K > 2$) or the various $K \times M$ weighted kappas for $K > 2$ or $M > 2$ in any context, either because they convey incomplete and possibly misleading information, or because other approaches are preferable to their use. We illustrate the use of the recommended kappas with applications in medical research. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: kappa; reliability; validity; consensus

1. INTRODUCTION

'Many human endeavors have been cursed with repeated failures before final success is achieved. The scaling of Mount Everest is one example. The discovery of the Northwest Passage is a second. The derivation of a correct standard error for kappa is a third'. This wry comment by Fleiss *et al.* in 1979 [1] continues to characterize the situation with regard to the kappas coefficients up to the year 2001, including not only derivation of correct standard errors, but also the formulation, interpretation and application of kappas.

*Correspondence to: Helena Chmura Kraemer, Department of Psychiatry and Behavioral Sciences, MC 5717, Stanford University, Stanford, CA 94305, U.S.A.
†E-mail: hck@leland.stanford.edu

Contract/grant sponsor: National Institute of Mental Health; contract/grant number: MH40041
Contract/grant sponsor: National Institute of Aging; contract/grant number: AG17824
Contract/grant sponsor: Department of Veterans Affairs Sierra-Pacific MIRECC
Contract/grant sponsor: Medical Research Service of the Department of Veterans Affairs

Received April 2001
Accepted October 2001

The various kappa coefficients are measures of association or correlation between variables measured at the categorical level. The first formal introductions of kappa were those, more than 40 years ago, by Scott [2] and Cohen [3]. Since then, the types of research questions in medical research that are well addressed with kappas (for example, reliability and validity of diagnosis, risk factor estimation) abound, and such areas of research have become of ever growing interest and importance [4]. Not surprisingly, numerous papers both using and criticizing the various forms of kappas have appeared in the statistical literature, as well as in the psychology, education, epidemiology, psychiatry and other medical literature. It is thus appropriate, despite the many existing 'revisits' of kappas [5–15], to take stock of what kappas are, what they are well-designed or ill-designed to do, and to bring up to date where kappas stand with regard to their applications in medical research.

To set the stage for discussion let us consider five major issues concerning kappas that are often forgotten or misinterpreted in the literature:

1. *Kappa has meaning beyond percentage agreement corrected for chance* (*PACC*). Sir Alexander Fleming in 1928 discovered penicillin by noticing that bacteria failed to grow on a mouldy Petri dish. However, in summarizing current knowledge of penicillin and its uses, a mouldy Petri dish is at most a historical curiosity, not of current relevance to knowledge about penicillin. In much the same way, Jacob Cohen discovered kappa by noticing that this statistic represented percentage agreement between categories corrected for chance (PACC). Since then, there has also been much expansion and refinement of our knowledge about kappa, its meaning and its use. Whether to use or not use kappa has very little to do with its relationship to PACC. With regard to kappa, that relationship is a historical curiosity. Just as some scientists study moulds, and others bacteria, to whom penicillin is a side issue, there are scientists specifically interested in percentage agreement. To them whether rescaling it to a kappa is appropriate to its understanding and use is a side issue [16–20]. Consequently there are now two separate and distinct lines of inquiry, sharing historical roots, one concerning use and interpretation of percentage agreement that will not be addressed here, and that concerning use and interpretation of kappa which is here the focus.
2. *Kappas were designed to measure correlation between nominal*, *not ordinal*, *measures*. While the kappas that emerged from consideration of agreement between non-ordered categories can be extended to ordinal measures [21–23], there are better alternatives to kappas for ordered categories. Technically, one can certainly compute kappas with ordered categories, for example, certain, probable, possible and doubtful diagnosis of multiple sclerosis [24], and the documentation of many statistical computer programs (for example, SAS) seem to support this approach, but the interpretation of the results can be misleading. In all that follows, the measures to be considered will be strictly nominal, not ordered categories.
3. *Even restricted to non-ordered categories*, *kappas are meant to be used*, *not only as descriptive statistics*, *but as a basis of statistical inference*. RBI or batting averages in baseball are purely descriptive statistics, not meant to be used as a basis of statistical inference. Once one understands how each is computed, it is a matter of personal preference and subjective judgement which statistic would be preferable in evaluating the performance of batters. In contrast, means, variance, correlation coefficients etc., as they are used in medical research, are descriptive statistics of what is seen in a particular

sample, but are also meant to estimate certain clinically meaningful population characteristics, and to be used as a basis of inference from the sample to its population. To be of value to medical research, kappas must do likewise.

Nevertheless, presentations of kappas often do not define any population or any parameter of the population that sample kappas are meant to estimate, and treat kappas purely as descriptive statistics [7]. Then discussions of bias, standard error, or any other such statistical inference procedures from sample to population are compromised. Many of the criticisms of kappas have been based on subjective opinions as to whether kappas are 'fair to the raters' or 'large enough', behave 'as they should', or accord with some personal preference as to what 'chance' means [7, 13, 25, 26]. These kinds of discussions of subjective preferences are appropriate to discussing RBI versus batting average, but not to estimation of a well-defined parameter in a population. We would urge that the sequence of events leading to use of a kappa coefficient should be: (i) to start with an important problem in medical research; (ii) to define the population and the parameter that the problem connotes; (iii) to discuss how (or whether) sample kappa might estimate that parameter, and (iv) to derive its statistical properties in that population. When this procedure is followed, it becomes clear that there is not one kappa coefficient, but many, and that which kappa coefficient is used in which situation is of importance. Moreover, there are many situations in which kappa can be used, but probably should not be.

4. *In using kappas as a basis of statistical inference, whether or not kappas are consistent with random decision making is usually of minimal importance.* Tests of the null hypothesis of randomness (for example, chi-square contingency table analyses) are well established and do not require kappa coefficients for implementation. Kappas are designed as effect sizes indicating the degree or strength of association. Thus bias of the sample kappas (relative to their population values), their standard errors (in non-random conditions), computation of confidence intervals, tests of homogeneity etc. are the statistical issues of importance [27–30]. However, because of overemphasis on testing null hypotheses of randomness, much of the kappa literature that deals with statistical inference focuses not on kappa as an effect size, but on testing whether kappas are random or not. In this discussion no particular emphasis will be placed on the properties of kappas under the assumption of randomness.
5. *The use of kappas in statistical inference does not depend on any distributional assumptions on the process underlying the generation of the classifications.* However, many presentations impose such restricting assumptions on the distributions of $\mathbf{p}_i$ that may not well represent what is actually occurring in the population.

The population model for a nominal rating is as follows. Patients in a population are indexed by i, $i=1,2,3,\ldots$. A single rating of a patient is a classification of patient i into one of $K(K>1)$ mutually exclusive and exhaustive non-ordered categories and is represented by a K-dimensional vector $\mathbf{X}_i=(X_{i1},X_{i2},\ldots,X_{iK})$, where $X_{ij}=1$, if patient i is classified into category j, and all other entries equal 0. For each patient, there might be M $(M>1)$ such ratings, each done blinded to all the others. Thus any correlation between the ratings arises from correlation within the patients and not because of the influence of one rater or rating on another. The probability that patient i $(i=1,2,\ldots)$ is classified into category j $(j=1,2,\ldots,K)$ is denoted p_{ij}, and $\mathbf{p}_i$ is the K-dimensional vector $(p_{i1},p_{i2},\ldots,p_{iK})$ with non-negative entries summing to 1. In a particular population of which patient i is a member, $\mathbf{p}_i$ has some, usually unknown, distribution over the $K-1$ dimensional unit cube.

For example, when there are two categories ($K=2$), for example, diagnosis of disease positive or negative, one common assumption is that the probability that a patient actually has the disease is π, and that if s/he has the disease, there is a fixed probability of a positive diagnosis ($X_{i1}=1$), the *sensitivity* (Se) of the diagnosis ($p_{i1}=\mathrm{Se}$); if s/he does not have the disease ($X_{i2}=2$), a fixed probability of a negative diagnosis, the *specificity* (Sp) of the diagnosis ($1-p_{i1}=\mathrm{Sp}$). This limits the distribution of p_{i1} to two points, Se and $1-\mathrm{Sp}$ ($p_{i2}=1-p_{i1}$): the 'sensitivity/specificity model' [31].

In the same situation, another model suggested has been the 'know/guess' model [25, 32, 33]. In this case, it is assumed that with a certain probability, π_1, a patient will be known with certainty to have the disease ($p_{i1}=1$); with a certain probability, π_0, a patient will be known with certainty not to have the disease ($p_{i1}=0$). For these patients, there is no probability of classification error. Finally, with the remaining probability, $1-\pi_1-\pi_0$, the diagnosis will be guessed with probability $p_{i1}=\alpha$. This limits the distribution of p_{i1} to 3 points $(1,\alpha,0)$.

One can check the fit of any such model by obtaining multiple blinded replicate diagnoses per patient. For these two models, three blinded diagnoses per patient would be required to estimate the three parameters in each model, $(\pi,\mathrm{Se},\mathrm{Sp})$ or (π_1,π_0,α), and at least one additional diagnosis per patient to test the fit of the model. In practice, it is hard to obtain four or more diagnoses per patient for a large enough sample size for adequate power, but in the rare cases where this has been done, such restrictive models are often shown to fit the data poorly [34]. If inferences are based on such limiting distributional assumptions that do not hold in the population, no matter how reasonable those assumptions might seem, or how much they simplify the mathematics, the conclusions drawn on that basis may be misleading. Kappas are based on no such limiting assumptions. Such models merely represent special cases often useful for illustrating certain properties of kappa, or for disproving certain general statements regarding kappa, as they here will be.

2. ASSESSMENT OF RELIABILITY OF NOMINAL DATA: THE INTRACLASS KAPPA

The reliability of a measure, as technically defined, is the ratio of the variance of the 'true' scores to that of the observed scores, where the 'true' score is the mean over independent replications of the measure [35, 36]. Since the reliability of a measure, so defined, indicates how reproducible that measure will be, how attenuated correlations against that measure will be, what loss of power of statistical tests use of that measure will cause, as well as how much error will be introduced into clinical decision making based on that measure [37], this is an important component of the quality of a measure both for research and clinical use. Since one cannot have a valid measure unless the measure has some degree of reliability, demonstration of reliability is viewed as a necessary first step to establishing the quality of a measure [14, 38].

The simplest way to estimate the reliability of a measure is to obtain a representative sample of N patients from the population to which results are to be generalized. (The same measure may have different reliabilities in different populations.) Then M ratings are sampled from the finite or infinite population of ratings/raters to which results are to be generalized, each obtained blinded to every other. Thus the ratings might be M ratings by the same pathologist of tissue slides presented over a period of time in a way that ensures blindness: *intra-observer reliability*. The ratings might be diagnoses by M randomly selected clinicians from a pool of clinicians all observing the patient at one point in time: *inter-observer reliability*. The ratings might be observations by randomly selected observers from a pool of observers, each observing the patient at one of M randomly selected time points over a span of time during which the characteristic of the patient being rated is unlikely to change: *test–retest reliability*. Clearly there are many different types of reliability depending on when, by whom, and how the multiple blinded ratings for each patient are generated. What all these problems have in common is that because of the way ratings are generated, the M successive ratings per patient are 'interchangeable', that is, the process underlying the M successive ratings per patient has the same underlying distribution of $\mathbf{p}_i$, whatever that distribution might be [39].

2.1. The 2×2 intraclass kappa

The simplest and most common reliability assessment with nominal data is that of two ratings ($M=2$), with two categories ($K=2$). In that case, we can focus on the X_{i1} since $X_{i2}=1-X_{i1}$ and on p_{i1}, since $p_{i2}=1-p_{i1}$. Then $E(X_{i1})=p_{i1}$, the 'true score' for patient i, $E(p_{i1})=P, \text{variance}(p_{i1})=\sigma_p^2$. Thus by the classical definition of reliability, the reliability of X is variance(p_{i1})/variance(X_{i1})$=\sigma_p^2/PP'$, where $P'=1-P$.

This intraclass kappa, κ, may also be expressed as

$$\kappa=(p_0-p_c)/(1-p_c)$$

where p_0 is the probability of agreement, and $p_c=P^2+P'^2$, that is, the PACC, for this has been shown to equal σ_p^2/PP' [31]. So accustomed are researchers to estimating the reliability of ordinal or interval level measures with a product-moment, intraclass or rank correlation coefficient, that one frequently sees 'reliability' there *defined* by the correlation coefficient between test–retest data. In the same sense, for binary data the reliability coefficient is *defined* by the intraclass kappa.

The original introductions of kappa [3, 40] defined not the population parameter, κ, but the sample estimate k, where the probability of agreement is replaced by the observed proportion of agreement, and P is estimated by the proportion of the classifications that selected category 1. This was proposed as a measure of reliability long before it was demonstrated that it satisfied the classical definition of reliability [31]. Fortunately, the results were consistent. However, that sequence of events spawned part of the problems surrounding kappa, since it opened the door for others to propose various sample statistics as measures of binary reliability, without demonstration of the relationship of their proposed measure with reliability as technically defined. Unless such a statistic estimates the same population parameter as does the intraclass kappa, it is *not* an estimate of the reliability of a binary measure. However, there are other statistics when $M=2$, that estimate the same parameter in properly designed reliability studies (random sample from the population of subjects, and a random sample of blinded raters/ratings for each subject), such as all weighted kappas (not the same as an intraclass kappa as will be seen below), or the sample phi coefficient, the risk difference or the attributable risk. Typically these provide less efficient estimators than does the sample intraclass kappa.

It is useful to note that $\kappa = 0$ indicates either that the heterogeneity of the patients in the population is not well detected by the raters or ratings, or that the patients in the population are homogeneous. Consequently it is well known that it is very difficult to achieve high reliability of any measure (binary or not) in a very homogeneous population (P near 0 or 1 for binary measures). That is not a flaw in kappa [26] or any other measure of reliability, or a paradox. It merely reflects the fact that it is difficult to make clear distinctions between the patients in a population in which those distinctions are very rare or fine. In such populations, 'noise' quickly overwhelms the 'signals'.

2.2. The $K \times 2$ intraclass kappa

When there are more than two categories ($K > 2$) both $\mathbf{X}_i$ and $\mathbf{p}_i$ are K-dimensional vectors. The classical definition of reliability requires that the covariance matrix of $\mathbf{p}_i$, Σ_p, be compared with the covariance matrix of $\mathbf{X}_i$, Σ_X. The diagonal elements of Σ_p are $\kappa_j P_j P_j'$, where κ_j is the 2×2 intraclass kappa with category j versus 'not-j', a pooling of the remaining categories, P_j is the $E(p_{ij})$, $P_j' = 1 - P_j$. The off-diagonal elements are $\rho_{jj^*} P_j P_{j^*}$, $j \neq j^*$, with ρ_{jj^*} the correlation coefficient between p_{ij} and p_{ij^*}. The diagonal elements of Σ_X are $P_j P_j'$, and the off-diagonal elements are $-P_j P_{j^*}$.

What has been proposed as a measure of reliability is the $K \times 2$ intraclass kappa

$$\kappa = \text{trace}(\Sigma_p)/\text{trace}(\Sigma_X) = \Sigma(P_j P_j' \kappa_j)/\Sigma(P_j P_j')$$

Again it can be demonstrated that this is equivalent to PACC with p_0 again the probability of agreement, now with $p_c = \Sigma P_j P_j'$.

From the above, it is apparent that to obtain a non-zero $K \times 2$ intraclass kappa requires that only one of the K categories have non-zero κ_j. If that one category has reasonable heterogeneity in the population ($P_j P_j'$ large) and has large enough κ_j, the $K \times 2$ intraclass kappa may be large.

Consider the special case for $K = 3$, when $\mathbf{p}_i = (1, 0, 0)$ with probability π, and $\mathbf{p}_i = (0, 0.5, 0.5)$ with probability $\pi' = 1 - \pi$. In this case category 1 is completely discriminated from categories 2 and 3, but the decisions between 2 and 3 are made randomly. Then $\kappa_1 = 1$, and $\kappa_2 = \kappa_3 = \pi/(\pi + 1)$, and the 3×2 intraclass kappa is $3\pi/(3\pi + 1)$. When $\pi = 0.5$, for example, $\kappa = 0.60$, and $\kappa_2 = \kappa_3 = 0.33$, even if 2 and 3 are here randomly assigned. Such a large overall κ can be mistakenly interpreted as a good reliability for all three categories, where here clearly only category 1 is reliably measured.

No one index, the $K \times 2$ intraclass kappa or any other, clearly indicates the reliability of a multi-category X. For categorical data, one must consider not only how distinct each category is from the pooled remaining categories (as reflected in the κ_j, $j = 1, 2, \ldots, K$), but how easily each category can be confused with each other [13, 41]. Consequently, we would suggest that: (i) multi-category kappas are not used as a measure of reliability with $K > 2$ categories; (ii) that seeking any *single* measure of multi-category reliability is a vain effort; and (iii) at least the K individual category κ_j's be reported, but that, better yet, methods be further developed to evaluate the entire misclassification matrix [42]. In particular, the decision to recommend kappa with two categories, but to recommend against kappa with more than two categories, is not influenced by the fact that kappa is related to PACC in both cases.

Table I. Estimation of the $2\times M$ intraclass correlation coefficient in the Periyakoil *et al.* data, with s the number of positive (grief) classifications from the $M=4$ raters, f_s the proportion of items with that number, k_s the kappa coefficient based on omitting one subject with s positive classifications, and w_s the weight needed to calculate the asymptotic variance.

s	f_s	s/M	$1-s/M$	k_s
0	0.2029	0.0000	1.0000	0.5700
1	0.1739	0.2500	0.7500	0.5860
2	0.0870	0.5000	0.5000	0.5918
3	0.1159	0.7500	0.2500	0.5873
4	0.4203	1.0000	0.0000	0.5725

General formula for k: $k=1-M\Sigma f_s(sM)(1-sM)/((M-1)PP')$
$P=\Sigma f_s(sM)$

Jack-knife formulae: Jack-knife $k=Nk-(N-1)\text{average}(k_s)$
Jack-knife $\text{SE}^2=(N-1)^2 s_k^2/N$
$s_k^2=\text{sample variance}(k_s)$

Results from above case: $P=0.5942$
$k=0.5792$
Jack-knife $k=0.6429$

The 2×2 intraclass kappa seems ideal as a measure of binary reliability, but the $K\times 2$ intraclass kappa we recommend against as uninterpretable. What if one had only two categories, but $M>2$ raters?

2.3. The $2\times M$ (multi-rater) intraclass kappa

With only two categories, the reliability coefficient is still $\kappa=\sigma^2/PP'$, as shown above. The multi-rater sample kappa statistic [43] is based on comparing pairwise agreement among the $M(M-1)/2$ pairs of raters evaluating each patient with what would be expected if classifications were randomly made. This process has been shown to obtain the equivalent result as applying the formula for the intraclass ρ for interval data to these binary data [44]. This statistic estimates the same κ as does the 2×2 intraclass kappa. For a fixed sample size of subjects, the larger the M, the smaller the estimation error.

There are several ways to estimate intraclass kappa here, but the easiest both for theory and application requires that the data be organized by s, the number of positive (category 1) classifications per patient (See Table I, column 1), $s=0,1,2,\ldots,M$. The proportion of the N patients sampled who have s of the M categorizations positive is f_s. The formula for calculation is presented in Table I, along with a demonstration of the calculation of this statistic based on a study conducted by one of the authors (VSP).

In this case, $N=69$ items were sampled from the population of items that might be used to distinguish preparatory grief (category 1) from depression (category 2) in dying adult patients. The issue was to assess to what extent clinicians could reliably distinguish the two. Depression, when it exists, is hypothesized to diminish quality of the dying process but can be effectively treated, while normal preparatory grief, when it exists, is hypothesized to be a sign of positive coping with the dying process that should be facilitated. $M=4$ expert clinicians were sampled and complied with classifying each item as more indicative of preparatory grief or depression. The results appear in Table I, with $k=0.579$.

Table II. The population probability distribution of the number of positive responses with M raters, generated from the sensitivity/specificity model (model A: $\mathrm{Se}=0.60$, $\mathrm{Sp}=0.99$, $\pi=0.1525$) and the know/guess model (model B: $\pi_1=0.0250$, $\pi_0=0.7875$, $\alpha=0.4054$). Both models have $P=0.10$ and $\kappa=0.50$ to two decimal places. Implication: the distribution of responses for $M>2$ differ even when P and κ are the same.

Number of positive $=s$	$M=2$		$M=4$		$M=6$	
	A	B	A	B	A	B
0	85.5%	85.4%	81.8%	81.1%	79.8%	79.6%
1	9.0%	9.0%	5.6%	6.4%	5.4%	3.4%
2	5.5%	5.6%	5.3%	6.5%	2.2%	5.8%
3			5.3%	3.0%	4.2%	5.2%
4			2.0%	3.0%	4.7%	2.7%
5					2.8%	0.7%
6					0.7%	2.6%

While the standard error is known and easily accessible when $M=2$ [43, 45–47], to date when $M>2$ it is known and easily accessible only under the null hypothesis of randomness [43]. The calculation of the standard error in general when $M>2$ was described by Fleiss as 'too complicated for presentation' (reference [43], p. 232), referring readers to Landis and Koch [48]. Not only is this standard error difficult to access, but also it is not known exactly how accurate it is for small to moderate sample size. Part of the problem lies in attempting to obtain a general solution when there are more than two categories (where intraclass kappa may be misleading), and when the number of ratings per patient is itself a variable from patient to patient (which may be problematic). The situation with the $2\times M$ intraclass kappa is much simpler.

For patient i, with probability p_{i1}, the probability that s of the interchangeable independent M ratings will be positive is the binomial probability ($s=0,1,2,\ldots,M$) with probability p_{i1} the binomial probability (say $\mathrm{Bin}(s;\,p_{i1},M), s=0,1,2,\ldots,M$). The probability that a randomly sample subject will be positive is the expected value of $\mathrm{Bin}(s;\,p_{i1},M)$ over the unknown distribution of p_{i1}. This involves moments of the p_{i1} distribution up to order M. Since P and κ involve only the first two moments, the distribution of the number of positive responses is determined by P and κ only when $M=2$. Consequently the quest for a standard error of the $2\times M$ intraclass sample kappas for $M>2$ that involves only parameters P and κ, that is, only moments up to order 2, is one of those futile quests [49]. One might have many different distributions of p_{i1} that have the same first two moments (P and κ) but that differ in the higher moments. For each such distribution the sample distribution for the $2\times M$ intraclass sample kappa would differ. This fact differentiates the distribution theory of the intraclass kappa for binary data from that of the intraclass correlation coefficient, ρ, to which it is closely computationally related, for interchangeable normal variates, for in the latter case, the distribution is determined by ρ, however large the number of raters, M.

For example, in Table II, we present an example of a 'sensitivity/specificity' model and of a 'know/guess' model selected to have almost exactly the same $P=0.10$ and $\kappa=0.50$, and show the distribution of response for $M=2,4,6$. It can be seen that the population distributions are almost the same for $M=2$, slightly different for $M=4$ and very different for $M=6$. Thus, unless $M = 2$, one would not expect that the distributions of the $2 \times M$ intraclass kappa would be the same in these two cases, much less in all cases with $P = 0.10$ and $k = 0.50$.

The vector of observed frequencies of the numbers of positive responses has a multinomial distribution with probabilities determined by the expected values of $\text{Bin}(s{:}\,p_i, M)$. Thus one can use the methods derived by Fisher [50] to obtain an approximate (asymptotic) standard error of kappa. An approximate standard error of k can also be obtained very easily using jack-knife procedures omitting one patient at time [45, 47, 51–53], as shown in Table I. These results correspond closely to those derived in various ways for the 2×2 intraclass kappas [43, 46, 47, 54]. The jack-knife procedure is demonstrated in Table I. (As a 'rule of thumb', the minimum number of patients should exceed both $10/P$ and $10/P'$. When $P = 0.5$, 20 patients are minimal; when $P = 0.01$, no fewer than 1000 patients are needed.) A generalized version of the SAS program (SAS Institute Inc., Cary NC) that performs the calculations can be located at http://mirecc.stanford.edu

When there are a variable number of raters per patient, the problem becomes more complicated, since the exact distribution of responses changes as M varies, involving more or fewer unknown moments of the p_{i1} distribution. If the patient's number of ratings is totally independent of his/her p_{i1}, one could stratify the patients by the number of ratings, obtain a 2×2 intraclass kappa from those with $M = 2$, a 2×3 intraclass kappa from those with $M = 3$ etc., and a standard error for each. Since these are independent samples from the same parent population, one could then obtain a weighted average of the kappas and its standard error using standard methods.

However, often the variation of the number of ratings is related to p_{i1}. Patients with more serious illnesses, for example, are more likely to have a positive diagnosis and less likely to provide the greater number of ratings. In that case, the subsamples of patients with $2, 3, 4, \ldots$ ratings may represent different populations and thus have different reliabilities that should not be muddled. This raises some serious questions about the practical application of the standard error derived by Landis and Koch [48] or any solution in which the number of ratings is variable.

To summarize, for the purpose of measuring reliability of a binary measure, the $2 \times M$ ($M \geqslant 2$) is highly recommended, but the use of the $K \times M$ kappa for $K > 2$ is questionable. To this it should be added that useful standards have been suggested for evaluation of the $2 \times M$ kappa as a measure of reliability [24], with $k \leqslant 0.2$ considered slight, $0.2 < k \leqslant 0.4$ as fair; $0.4 < k \leqslant 0.6$ as moderate, $0.6 < k \leqslant 0.8$ as substantial and $k > 0.8$ as almost perfect. It is important to realize that a kappa coefficient below 0.2 is slight, no matter what the p-value is of a test of the null hypothesis of randomness. Moreover, a kappa coefficient above 0.6 that is not 'statistically significant' on such a test indicates inadequate sample size, not a definitive conclusion about the reliability of the measure. It is the magnitude of k that matters, and how precisely that is estimated, not the p-value of a test of the null hypothesis of randomness [55].

3. VALIDITY OF CATEGORICAL MEASURES: THE $K \times M$ WEIGHTED KAPPAS

The validity of a measure is defined as the proportion of the observed variance that reflects variance in the construct the measure was intended to measure [36, 38], and is thus always no greater than the reliability of a measure. Validity is generally assessed by a correlation coefficient between a criterion or 'gold standard' (X_i) and the measure (Y_i) for each patient in a representative sample from the population to which the results are to be generalized. (Once again, a measure might be more valid in one population than in another.) If a measure is completely valid against a criterion, there should be a 1:1 mapping of the values of Y_i onto the values of X_i. With categorical measures, the hope is to be able to base clinical or research decisions on Y_i that would be the same as if those decisions were based on the 'gold standard' X_i. That would require not only that the number of categories of Y_i match the number of categories of X_i, but that the labels be the same.

The 'gold standard' is the major source of difficulty in assessing validity, for there are very few true 'gold standards' available. Instead, many 'more-or-less gold standards' are considered, each somewhat flawed, but each of which provides some degree of challenge to the validity of the measure. Thus, as in the case of reliability, there are many types of validity, depending on how the 'gold standard' is selected: face validity; convergent validity; discriminative validity; predictive validity; construct validity.

While there are many problems in medical research that follow this paradigm, few of which are actually labelled 'validity' studies, we will for the moment focus on medical test evaluation. In medical test evaluation, one has a 'gold standard' evaluation of the presence/absence or type of disease, usually the best possible determination currently in existence, against which a test is assessed. To be of clinical and policy importance the test result for each patient should correspond closely to the results of the 'gold standard', for treatment decisions for patients are to be based on that result.

3.1. A 2 × 2 weighted kappa coefficient

Once again the most common situation is with two ratings per patient, say X_i and Y_i each having only two categories of response. We use different designations here for the two ratings, X_i and Y_i, in order to emphasize that the decision process underlying the 'gold standard' (X_i) and the diagnosis under evaluation (Y_i) are, by definition, not the same. For the same reason, we focus on the probability of a positive result (category 1) in each case, with probability p_{i1} for X_i and q_{i1} for Y_i, using different notation for the probabilities.

The distribution of p_{i1} and q_{i1} in the population of patients may be totally different, even if $P = E(p_{i1})$ and $Q = E(q_{i1})$ are equal. The equality of P and Q cannot be used to justify the use of the intraclass kappa in this situation, for the intraclass kappa is appropriate only to the situation in which all the moments, not just the first, are equal (interchangeable variables).

Since X_i and Y_i are 'blinded' to each other, the probability that for patient i both X_i and Y_i are positive is $p_{i1}q_{i1}$. Thus in the population, the probability that a randomly selected patient has both X_i and Y_i positive is $E(p_{i1}q_{i1}) = PQ + \rho\sigma_p\sigma_q$, where $P = E(p_{i1})$, $Q = E(q_{i1})$, ρ is the product moment correlation coefficient between p_{i1} and q_{i1}, $\sigma_p^2 = \text{variance}(p_{i1})$, $\sigma_q^2 = \text{variance}(q_{i1})$. All the probabilities similarly computed are presented in Table III.

It can be seen in Table III that the association between X_i and Y_i becomes stronger as $\rho\sigma_p\sigma_q$ increases from zero. At zero, the results in the table are consistent with random decision making. Any function of $\rho\sigma_p\sigma_q$, P and Q, that is strictly monotonic in $\rho\sigma_p\sigma_q$, that takes on the value zero when $\rho = 0$, and takes on the value $+1$ when the probabilities on the cross diagonal are both 0, and -1 when the probabilities on the main diagonal are both 0, is a type of correlation coefficient between X and Y. The difficulty is that there are an infinite number of such functions (some of the most common defined in Table III), and therefore an

Table III. The 2×2 weighted kappa: probabilities and weights. Definitions of some common measures used in medical test evaluation or in risk assessment.

	$Y=1$	$Y=2$	Total
Probabilities			
$X=1$	$a=PQ+\rho\sigma_p\sigma_q$	$b=PQ'-\rho\sigma_p\sigma_q$	P
$X=2$	$c=P'Q-\rho\sigma_p\sigma_q$	$d=P'Q'+\rho\sigma_p\sigma_q$	$P'=1-P$
Total	Q	$Q'=1-Q$	
Weights indicating loss or regret $(0<r<1)$:			
$X=1$	0	r	
$X=2$	$r'=1-r$	0	

$\kappa(r)=(ad-bc)/(PQ'r+P'Qr')=\rho\sigma_p\sigma_q/(PQ'r+P'Qr'),(0<r<1)$.
$\kappa(1/2)=2(ad-bc)/(PQ'+P'Q)=(p_0-p_c)/(1-p_c),(p_0=a+d,p_c=PQ+P'Q')$.

Sensitivity of Y to X: $\mathrm{Se}=a/P=Q+Q'\kappa(1)$.
Specificity of Y to X: $\mathrm{Sp}=d/P'=Q'+Q\kappa(0)$.
Predictive value of a positive test: $\mathrm{PVP}=a/Q=P+P'\kappa(0)$.
Predictive value of a negative test: $\mathrm{PVN}=d/Q'=P'+P\kappa(1)$.
Percent agreement $=p_0=a+d=p_c+p_c'\kappa(1/2)$.
Risk difference $=\mathrm{Se}+\mathrm{Sp}-1=a/P-c/P'=\kappa(Q')$.
Attributable risk $=\kappa(0)$.
Odds ratio $=ad/bc=(\mathrm{SeSp})/(\mathrm{Se'Sp'})=(\mathrm{PVP\ PVN})/(\mathrm{PVP'\ PVN'})$.

infinite number of correlation coefficients that yield results not necessarily concordant with each other.

There is one such correlation coefficient, a certain 2×2 weighted kappa, unique because it is based on an acknowledgement that the *clinical* consequences of a false negative (X_i positive, Y_i negative) may be quite different from the *clinical* consequences of a false positive (X_i negative, Y_i positive) [47]. For example, a false negative medical test might delay or prevent a patient from obtaining needed treatment in timely fashion. If the test were to fail to detect the common cold, that might not matter a great deal, but if the test were to fail to detect a rapidly progressing cancer, that might be fatal. Similarly a false positive medical test may result in unnecessary treatment for the patient. If the treatment involved taking two aspirin and calling in the morning, that might not matter a great deal, but if it involved radiation, chemotherapy or surgical treatment, that might cause severe stress, pain, costs and possible iatragenic damage, even death, to the patient. The balance between the two types of errors shifts depending on the population, the disorder and the medical sequelae of a positive and negative test. This weighted kappa coefficient is unique among the many 2×2 correlation coefficients in that in each context of its use, it requires that this balance be explicitly assessed *a priori* and incorporated into the parameter.

For this particular weighted kappa, a weight indicating the clinical cost of each error is attributed to each outcome (see Table III); an index r is set that ranges from 0 to 1 indicating the relative importance of false negatives to false positives. When $r=1$, one is primarily concerned with false negatives (as with a screening test); when $r=0$, one is primarily concerned with false positives (as with a definitive test); when $r=1/2$, one is equally concerned with both (as with a discrimination test). The definition of $\kappa(r)$ in this case [47, 56] is

$$\kappa(r)=\rho\sigma_p\sigma_q/(PQ'r+P'Qr')$$

The sample estimator is $k(r) = (ad - bc)/(PQ'r + P'Qr')$, where a, b, c, d are the proportions of the sample in the cells so marked in Table III, P and Q estimated by the sample proportions. Cohen's kappa [40], often called the 'unweighted' kappa, is $\kappa(1/2)$

$$\kappa(1/2) = (p_0 - p_c)/(1 - p_c)$$

where p_0 again is the proportion of agreement, and here $p_c = PQ + P'Q'$, once again a PACC (see Table III for a summary of definitions). When papers or programs refer to 'the' kappa coefficient, they are almost inevitably referring to $\kappa(1/2)$, but it must be recognized that $\kappa(1/2)$ reflects a decision (conscious or unconscious) that false negatives and false positives are equally clinically undesirable, and $\kappa(r)$ equals PACC only when $r = 1/2$.

Different researchers are familiar with different measures of 2×2 association, and not all readers will be familiar with all the following. However, it is important to note the strong interrelationships among the many measures of 2×2 association. Risk difference (Youden's index) is $\kappa(Q')$, and attributable risk is $\kappa(0)$, reflecting quite different decisions about the relative importance of false positives and negatives. The phi coefficient is the geometric mean of $\kappa(0)$ and $\kappa(1)$: $(\kappa(0)\kappa(1))^{1/2}$. Sensitivity and predictive value of a negative test rescaled to equal 0 for random decision making and 1 when there are no errors, equal $\kappa(1)$. The specificity and predictive values of a positive test, similarly rescaled, equal $\kappa(0)$. For any r between 0 and 1, $\kappa(r)/\max\kappa(r)$ and phi/max phi [57], where max $\kappa(r)$ and max phi are the maximal achievable values of $\kappa(r)$ and phi, respectively, equal either $\kappa(0)$ or $\kappa(1)$, depending on whether P is greater or less than Q. This briefly demonstrates that most of the common measures of 2×2 association either (i) equal $\kappa(r)$ for some value of r, or, (ii) when rescaled, equal $\kappa(r)$ for some value of r, or (iii) equal some combination of the $\kappa(r)$. Odds ratio and measures of association closely related to odds ratio seem the notable exceptions.

Researchers sometimes see the necessity of deciding *a priori* on the relative clinical importance of false negatives versus false positives as a problem with $\kappa(r)$, since other measures of 2×2 association do not seem to require any such *a priori* declaration. In fact, the opposite is true. It has been demonstrated [58] that every measure of 2×2 association has implicit in its definition some weighting of the relative importance of false positives and false negatives, often unknown to the user. The unique value of this weighted kappa as a measure of validity is that it *explicitly* incorporates the relative importance of false positives and false negatives, whereas users of other 2×2 measures of association make that same choice by choosing one measure rather than another, and often do so unaware as to the choice they have *de facto* made. If they are unaware of the choice, that is indeed a problem, for there is risk of misleading clinical and policy decisions in the context in which the user applies it [58].

However, unlike the situation with reliability, it cannot be argued that $\kappa(r)$, in any sense, defines validity, for the appropriate choice of a validity measure depends on what the user stipulates as the relative importance of false positives and false negatives. How these are weighted may indicate a choice of index not directly related to any $\kappa(r)$ (the odds ratio, for example).

It is of importance to note how the relative clinical importance (r) and the reliabilities of X and Y (the intraclass κ_X and κ_Y defined above for X and Y) influence the magnitude of $\kappa(r)$:

$$\kappa(r) = \rho(\kappa_X \kappa_Y)^{1/2}(PP'QQ')^{1/2}/(PQ'r + P'Qr')$$

with $P' = 1 - P$, $Q' = 1 - Q$, $r' = 1 - r$.

Here, as defined above, ρ is the correlation between p_{i1} and q_{i1} (which does not change with r). κ_X and κ_Y are the test–retest reliabilities of X and Y (which do not depend on r). As is always expected of a properly defined reliability coefficient, the correlation between X and Y reflected in $\kappa(r)$ suffers attenuation due to the unreliabilities of X and Y, here measured by the intraclass kappas κ_X and κ_Y. Only the relationship between P and Q affects $\kappa(r)$ differently for different values of r. When $P=Q$, $\kappa(r)$ is the same for all values of r and estimates the same population parameter as does the intraclass kappa although the distribution of the sample intraclass kappa is not exactly the same as that of the sample weighted kappa. For that matter, when $P=Q$, the sample distributions of $k(r)$ for different values of r are not all the same, even though all estimate the same parameter. Otherwise, in effect, too many positive tests ($Q>P$) are penalized by $\kappa(r)$ when false positives are of more concern (r nearer 0), and too many negative tests ($Q<P$) are penalized by $\kappa(r)$ when false negatives are of more concern (r nearer 1).

A major source of confusion in the statistical literature related to kappa is the assignment of weights [13]. Here we have chosen to use weights that indicate loss or regret, with zero loss for agreements. Fleiss [43] used weights that indicate gain or benefit, with maximal weights of 1 for agreements. Here we propose that false positives and false negatives may have different weights. Fleiss required that they be the same. Both approaches are viable for different medical research problems, as indeed are many other sets of weights, including sets that assign different weights to the two types of agreements.

If the weights reflect losses or regrets, $\kappa(r)=(E_c(r)-E_o(r))/(E_c(r)-\min)$, while if the weights reflect gains or benefits, $\kappa(r)=(E_o(r)-E_c(r))/(\max-E_c(r))$, where $E_c(r)$ is the expected weight when $\rho=0$ and $E_o(r)$ the expected weight with the observed probabilities. The scaling factor min is the ideal minimal value of $E_o(r)$ when losses are considered, and max is the ideal maximal value of $E_o(r)$ when gains are considered, for the particular research question. Here min is 0, where there are no disagreements; Fleiss' max is 1, also when there are no disagreements. Regardless of the weight assigned to disagreements in Fleiss' version of kappa, his weighted kappas in the 2×2 situation all correspond to what is here defined as $\kappa(1/2)$, while if P and Q are unequal, here $\kappa(r)$ changes with r, and generally equals $\kappa(1/2)$ only when $r=1/2$.

How the weights, min and max, are assigned changes the sampling distribution of $\kappa(r)$, which may be one of the reasons finding its correct standard error has been so problematic. Since the weights should be dictated by the nature of the medical research question, they should and will change from one situation to another. It is not possible to present a formula for the standard error that would be correct for all possible future formulations of the weights. For the particular weights used here (Table III) the Fisher procedure [50] could be used to obtain an asymptotic standard error. However, given the difficulties engendered by the wide choice of weights, and the fact that it is both easier and apparently about as accurate [54] when sample size is adequate, we would here recommend instead that the jack-knife estimator be used. That would guarantee that the estimate of the standard error be accurate for the specific set of weights selected and avoid further errors.

3.2. *The $K\times 2$ multi-category kappa*

In the validity context, as noted above, if the 'gold standard' has K categories, any candidate valid measure must also have K categories with the same labels. Thus, for example,

Table IV. Example: the joint probability distribution of a three-category X and a three-category Y, with one perfectly valid category ($Y=1$ for $X=1$), and two invalid categories ($Y=2$ for $X=2$) and ($Y=3$ for $X=3$) because of an interchange of $Y=2$ and $Y=3$ ($P_1+P_2+P_3=1$).

	$Y=1$	$Y=2$	$Y=3$	Total
$X=1$	P_1	0	0	P_1
$X=2$	0	0	P_2	P_2
$X=3$	0	P_3	0	P_3
Total	P_1	P_3	P_2	

if the 'gold standard' identifies patients with schizophrenia, depression, personality disorder, and 'other', any potentially valid diagnostic test would also identify the same four categories. In a direct generalization of the above, if 'gold standard' and diagnosis agree, disagreement is zero. If, however, someone who is schizophrenic is treated for depression, that is not an error necessarily of equal clinical importance as someone who is depressed being treated for schizophrenia. For each possible disgreement, one could assess the relative clinical importance of that misclassification, denoted r_{jj*} for $j \neq j^*$. The only requirement is that $r_{jj*} \geqslant 0$ for all $j \neq j^*$, and that $\Sigma r_{jj*}=1$. Then the weighted kappa, $\kappa(r)$, is defined as above as $(E_c(r)-E_o(r))/E_c(r)$.

The difficulty here, as with the $K \times 2$ intraclass kappa, is that $\kappa(r)$ is sure to equal 0 only if *all* classifications are random. Thus having only one valid category can yield a positive $\kappa(r)$, or we might have $\kappa(r)$ near zero when all but one category are completely valid.

For example, consider the case shown in Table IV. Here diagnostic category 1 is completely valid for 'gold standard' category 1, but diagnostic categories 2 and 3 are obviously switched. When (all r_{jj*} here equal) $P_1=0.1$, $P_2=0.4$ and $P_3=0.5$, $k(r)=-0.525$. When $P_1=0.3$, $P_2=0.5$, $P_3=0.2$, $k(r)=+0.014$. When $P_1=0.8$, $P_2=P_3=0.1$, $k(r)=+0.412$. None of these results $(-0.525,+0.014,+0.412)$ suggests what is obvious from examination of the complete cross-classification matrix: Y-categories 2 and 3 must be switched to obtain perfect validity. Consequently, once again, we propose that, like the multi-category intraclass kappa, the multi-category weighted kappas not be used as a measure of validity, for no single measure of validity can convey completely and accurately the validity of a multi-category system, where some categories may be valid but vary in terms of degree of validity, and others may be invalid.

3.3. The $2 \times M$ Multi-rater kappa

Now suppose that we had a binary 'gold standard' X_i, and M binary diagnostic tests: $Y_{i1}, Y_{i2}, \ldots, Y_{iM}$. Can the M diagnostic tests be used to obtain a valid diagnosis of X_i, and how valid would that test be? In this case, X_i and each Y_{ij} may have a different underlying distribution of p_{i1} or q_{i1}. While we could propose a multi-rater kappa [59], generally the way this problem is approached in medical test evaluation is by developing a function $g(Y_{i1}, Y_{i2}, \ldots)$, called a 'risk score', such that $g()$ is monotonically related to $\mathrm{Prob}(X_i=1)$. Then some cutpoint is selected so that if $g(Y_{i1}, Y_{i2}, \ldots) \geqslant C$, the diagnostic test is positive, and otherwise negative.

Almost inevitably, applying such a cutpoint dichotomizing the ordinal risk score to a binary classification reduces the power of statistical tests based on the measures [60]. If the cutpoint is injudiciously chosen, it may also mislead research conclusions. However, for clinical decision making, that is, deciding who to treat and not treat for a condition, who to hospitalize or not, a binary measure is necessary. Thus while the recommendation not to dichotomize for purposes of research is almost universal, dichotomization for clinical purposes is often necessary. Such dichotomization reduces the multivariate tests to a binary test based on all the individual tests. The 2×2 weighted kappa may then be used as a measure of the validity of the combined test.

The most common method of developing this function is multiple logistic regression analysis where it is assumed that logit $\text{Prob}(X_i = 1|Y_{i1}, Y_{i2}, \ldots) = \beta_0 + \Sigma\beta_j Y_{ij}$, that is, some linear function of the Y's, with a 'risk score' $(\Sigma\beta_j Y_{ij})$ assigned to each patient. Regression trees [56, 61] can also be used, using whatever validity criterion the developer chooses to determine the optimal test at each stage and a variety of stopping rules. Each patient in a final branch is given a 'risk score' equal to the $\text{Prob}(X_i = 1)$ in that subgroup. Finally, one might simply count the number of positive tests for each patient, $g(Y) = \Sigma Y_{ij}$, and use this as a 'risk score'. There are many such approaches, all of which reduce the 2^M possible different responses to the M binary tests to a single ordinal response, the 'risk score', using all M tests in some sense optimally. The relative strengths and weaknesses of these and other approaches to developing the 'risk score' can be vigorously debated. However, that is not the issue here.

When the 'risk score' is determined, the cutpoint C is often selected to equate P and Q, that is, so that $Q = \text{Prob}(g(Y_{i1}, Y_{i2}, \ldots) \geqslant C) = P$. This is not always ideal. Better yet, the optimal cutpoint would be the one that maximizes $\kappa(r)$, where r again indicates the relative importance of false negatives to false positives [56], or whichever other measure of 2×2 association best reflects the trade-offs between false positives and false negatives.

We do not recommend any $2\times M$ weighted kappa coefficient as a measure of validity, for there are already a variety of other standard methods used in this problem that seem to deal well with the problem. None seems to require or would benefit from a $2\times M$ kappa coefficient, for all focus more appropriately on reducing the problem to a 2×2 problem. Then the 2×2 weighted kappa might be used as a measure of validity.

4. THE PROBLEM OF CONSENSUS DIAGNOSIS

The final context of medical research in which kappa coefficients have proved uniquely useful is that of the consensus diagnosis. Suppose one assesses the reliability of a binary X_i, and found that its reliability, as measured by a $2\times M$ intraclass kappa, was greater than zero, but not satisfactory. 'Rule of thumb' standards for reliability have been proposed [14, 24]. By those standards, $\kappa = 0.579$, as in the Periyakoil data, or $\kappa = 0.5$, as in both cases of Table II, would be considered 'moderate' [24] or 'fair' [14]. Could one use a consensus of M raters, requiring at least C positive diagnoses for a consensus positive diagnosis, and thereby achieve adequate (say $\kappa > 0.8$, 'almost perfect' or 'substantial') reliability? How large should M be, and what value of C should be chosen?

One could deal with the problem using brute force: sample $2M$ raters for each patient sampled, randomly split the raters into two groups of M for each patient. Then for $C = 1, 2, \ldots, M$, determine the diagnosis for that value of C, and obtain 2×2 intraclass kappa, κ_{CM}. Then choose the optimal cutpoint C as the one that maximizes κ_{CM} for that value of M. Then vary M.

Table V. The optimal consensus diagnoses for the sensitivity/specificity model with $\mathrm{Se}=0.60$. $\mathrm{Sp}'=0.01$, $\pi=0.1525$, and for the know/guess model with $\pi_1=0.0250$, $\pi_0=0.7875$, $\alpha=0.4054$. Both models have $P=0.10$, $\kappa=0.50$. The number of diagnoses in the consensus is M, with C the optimal cutpoint (a positive diagnosis is given those with C or more positive diagnoses of the M). Q is the proportion diagnosed positive with the optimal consensus, and κ is the intraclass κ for that consensus.

M	Sensitivity/specificity model			Know/guess model		
	C	Q	κ	C	Q	κ
1	1	0.10	0.50	1	0.10	0.50
2	1	0.14	0.70	1	0.15	0.66
3	1	0.17	0.76	1	0.17	0.78
4	2	0.13	0.79	1	0.19	0.87
5	2	0.14	0.89	1	0.20	0.92
6	2	0.15	0.94	1	0.20	0.95
7	2	0.15	0.96	1	0.21	0.97
8	2	0.15	0.97	1	0.21	0.98
9	2	0.15	0.97	1	0.21	0.99
10	3	0.15	0.98	1	0.21	0.99

With the four raters in Table I, we have already calculated that $\kappa_{11}=0.579$. We then randomly split the pool of four raters into two sets of two for each patient, and found that $\kappa_{12}=0.549$, and $\kappa_{22}=0.739$. Thus the optimal consensus of 2 is to use a cutpoint $C=2$, and the reliability then rises from $\kappa_{11}=0.579$ with one rater to $\kappa_{22}=0.739$ for an optimal consensus of two. For an expanded discussion of these methods, see Noda *et al.* [62], and for a program to perform such calculations see http://mirecc.stanford.edu

It is of note that if the optimal consensus of 2 is obtained when $C=1$, in practice one would not request a second opinion when the first one was positive. If, as above, the optimal consensus of 2 is obtained when $C=2$, in practice one would not request a second opinion when the first one was negative. It often happens with the optimal consensus that, when put into practice, the number of ratings per patient to obtain a consensus of M is far less than M ratings per patient. This often means one can increase the quality of the diagnosis with minimal increase in time and cost. However, to identify that optimal consensus in the first place requires $2M$ ratings for each patient. Thus to evaluate a consensus of 3, one needs 6 ratings per patient, for 4, one needs 8, etc. This rapidly becomes an unfeasible solution in practice.

The theoretical solution is easy. For a patient with probability p_{i1} on a single rating, the probability of a positive diagnosis for a consensus of C of M is

$$q_{iCM}=\mathrm{Bin}(C;\,p_{i1},M)$$

where $\mathrm{Bin}(C;\,p_{i1},M)$ is the probability that a binomial random variable with parameters p_i and M equals or exceeds C. Thus $Q_{CM}=E(\mathrm{Bin}(C;\,p_{i1},M))$ and $\kappa_{CM}=\mathrm{var}(q_{iCM})/(Q_{CM}Q'_{CM})$. If we knew the distribution of p_{i1}, we would also know the distribution of q_{iCM} for all C and M, and thus know κ_{CM}. In Table V, for example, are presented the two hypothetical cases of Table II, where we do know the distribution and they have almost identical P and κ. Here for the sensitivity/specificity model, as M increases from 1 to 10, the optimal C rises from 1 to 2 to 3, and the κ from 0.50 for one observation to 0.98 for a consensus of 10. One would need a consensus of 2 positive out of 5 to achieve $\kappa \geqslant 0.8$. On the other hand, for the 'know/guess' model, as M increases from 1 to 10, the optimal C is always equal to 1, but the κ still rises from 0.50 for one observation to 0.99 for a consensus of 10. One would now need a consensus of 1 positive out of 4 to achieve $\kappa \geqslant 0.8$.

The above illustration demonstrates the fallacy of certain intuitive notions:

(i) It is not necessarily true that the optimal consensus equates Q with P.
(ii) The 'majority rule' (always use the first C exceeding $M/2$), is not always best.
(iii) The 'unanimity rule' (always use $C=0$ or $C=M$), too, is not always best.
(iv) Knowing P and κ does not settle the issue, for quite different optimal consensus rules were derived for the two situations in Table II having almost the same P and κ.

Since the reason for dichotomization is most compelling for purposes of clinical decision making, these false intuitive notions can mislead such decisions.

Examination of cases such as these provides some insight into the solution. For the 'sensitivity/specificity' model, it can be seen that for every M, the optimal C cuts off as close to the top 15 per cent of the number of positives as is possible. That 15 per cent corresponds to the 'high risk' subgroup with $p_{i1} = \mathrm{Se} = 0.60$. For the 'know/guess' model, the optimal C cuts off as close to the top 21 per cent of the number of positives as is possible. That 21 per cent corresponds to the 'high risk' comprising the subgroup of 2.5 per cent with $p_{i1} = 1$ plus the subgroup of 18.8 per cent with $p_{i1} = 0.4054$. However, in general, what proportion Q^* constitutes the 'high risk' subgroup?

The numerator of κ is $\mathrm{var}(p_{i1})$ which, for any P^* between 0 and 1, can be partitioned into two components:

$$\mathrm{var}(p_{i1}) = 2Q^*Q^{*\prime}(\mu_1 - \mu_2)^2 + Q^*\mathrm{var}(p_{i1} \mid p_{i1} \geqslant P^*) + Q^{*\prime}\mathrm{var}(p_{i1} \mid p_{i1} < P^*)$$

where $Q^* = \mathrm{prob}(p_{i1} \geqslant P^*)$, $Q^{*\prime} = 1 - Q^*$, $\mu_1 = E(p_{i1} \mid p_{i1} \geqslant P^*)$, and $\mu_2 = E(p_{i1} \mid p_{i1} < P^*)$. The percentage cut off by optimal C approximates Q^*, for that value of P^* for which the first term of $\mathrm{var}(p_{i1})$ is maximized. Thus the optimal cutpoint for p_{i1} (P^*), which determines the percentage of 'high risk' subjects (Q^*), is determined by what dichotomization of the p_{i1} distribution absorbs as much of the variance as possible [63].

5. CONCLUSIONS

To summarize:

(i) The $2 \times M$ intraclass kappa ($M \geqslant 2$) for a well-designed reliability study directly estimates reliability as defined in the classical sense and is thus the ideal reliability coefficient for a binary measure. For reasonable sample size, its standard error can be easily computed, and used to formulate confidence intervals, to test homogeneity of κ's and to address other such statistical challenges, such as developing optimal consensus rules.
(ii) The 2×2 weighted kappa $\kappa(r)$ described here is an excellent choice as a validity measure, although not a unique choice. However, since it explicitly requires that the relative importance of false positives and false negatives be specified and incorporated

into the validity measure, while all other 2×2 measures require that choice implicitly, $\kappa(r)$ is highly recommended in this context. For reasonable sample size, its standard error can easily be computed using jack-knife methods.

(iii) The $K \times M$ intraclass kappa, for $K > 2$, is not recommended as a measure of reliability, for no single measure is sufficient to completely and accurately convey information on reliability when there are more than two categories.

(iv) The $K \times M$ weighted kappas for $K > 2$ or $M > 2$ are not recommended as validity measures. When $K > 2$, the situation is similar to that with the $K \times M$ intraclass kappa. Any single measure, including $\kappa(r)$, is not enough to provide the necessary information on validity when some categories may be valid and others not. When $M > 2$, all the preferred methods in one way or another dichotomize the multi-dimensional Y space to create a binary outcome, and may then choose to use the 2×2 weighted kappa as a measure of validity. A $K \times M$ weighted kappa is not needed.

Even limited to these two contexts of reliability and validity, a broad spectrum of important medical research problems are encompassed. The $2 \times M$ intraclass kappa applies to any situation in which units are sampled from some population, and multiple subunits are sampled from each unit, where the intra-unit concordance or the inter-unit heterogeneity is of research interest.

For example, the intraclass kappa is useful as a measure of twin concordance in genetic studies of twins [64] (and could be used for triplets or quadruplets), as a measure of inter-sibling concordance in family studies, of intra-group concordance among patients in a therapy group etc. A research question such as the following also falls into the same category: If one sampled physicians or hospitals who performed a certain procedure, and assessed the outcome (success/failure) on a random sample of M of each physician's or hospital's patients undergoing that procedure, how heterogeneous would the physicians or hospitals prove to be? Here $\kappa = 0$ would indicate absolute homogeneity of results; larger κ would indicate greater heterogeneity (perhaps related to the type of patients referred, training, skill, resources or experience). Moreover, if there were a hypothesized source of heterogeneity (perhaps those that specialize in that procedure versus those that only occasionally do it), one could stratify the population by that source, compute the $2 \times M$ intraclass kappa within each stratum. If indeed that source accounted for most of the heterogeneity, the $2 \times M$ intraclass kappa within each stratum would approach zero.

The 2×2 weighted kappa in general could be applied to any situation in which the correlation between binary X_i and binary Y_i is of interest, where there are clinical consequences to be associated with the decisions. The particular weighted kappa discussed here is particularly relevant when Y_i is to be used to make decisions relative to X_i, in which case it is prudent to consider the relative clinical importance of false positives and false negatives. There are a vast number of research questions of this type in medical research. We have used as an example the evaluation of a medical test against a binary 'gold standard'. Since such medical tests are often the basis of medical decisions of whom to treat and how, such problems are of crucial importance. However, X_i might also represent the presence or absence of a disorder, and Y_i a possible risk factor for that disorder. Such information often influences policy recommendations as to preventive measures or targeting of certain populations for preventive interventions. In that situation, $\kappa(r)$ would be used as a measure of potency of that risk factor [58]. X_i might be the diagnosis by an acknowledged expert, and Y_i the diagnosis by a less expert clinician, a nurse, a layman, from a different source, or under different conditions. Such questions often arise in health services research, for if one can achieve the same (or better) quality of diagnosis from less costly sources, one could decrease medical costs with no decrease in quality of care. What characterizes all these situations is that X_i is a criterion against which Y_i is to be evaluated, and that there are costs to misclassification that are embodied in the weight, r, that defines $\kappa(r)$.

There are many other medical research questions for which some form of kappa could conceivably be used, but to date, the logic of suggesting any form of kappa is either absent or weak. For example, to show that two disorders are non-randomly comorbid in a population, one would assess how frequently they co-occur in that population and show this is more frequent than random association would suggest [65]. One could certainly use a kappa to measure such comorbidity, but which kappa, and why any kappa would be preferable to the odds ratio, for example, is not clear. If one were interested in whether one could use a single nominal observation plus other information to predict a second nominal observation, one might prefer various regression modelling approaches, such as log-linear models [5, 20, 66]. So far, there appear to be few other contexts not covered above, where use of a kappa coefficient might be unequivocally recommended or preferred to other methods. Thus it appears that there are certain situations where kappa coefficients are ideally suited to address research questions ($2 \times M$ intraclass kappa for reliability), certain situations in which kappa coefficients have qualities that make them outstanding choices (2×2 weighted kappa in the validity context), and many other situations in which kappa coefficients may mislead or where other approaches might be preferable.

ACKNOWLEDGEMENTS

This work was supported in part by the National Institute of Mental Health grant MH40041, the National Institute of Aging grant AG17824, the Department of Veterans Affairs Sierra-Pacific MIRECC, and the Medical Research Service of the Department of Veterans Affairs.

REFERENCES

1. Fleiss JL, Nee JCM, Landis JR. Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin* 1979; **86**:974–977.
2. Scott WA. Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly* 1955; 321–325.
3. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960; **20**:37–46.
4. Feinstein AR. A bibliography of publications on observer variability. *Journal of Chronic Diseases* 1985; **38**: 619–632.
5. Banerjee M, Capozzoli M, McSweeney L, Sinha D. Beyond kappa: a review of interrater agreement measures. *Canadian Journal of Statistics* 1999; **27**:3–23.
6. Bartko JJ, Carpenter WT. On the methods and theory of reliability. *Journal of Nervous and Mental Disease* 1976; **163**:307–317.
7. Brennan RL, Prediger DJ. Coefficient kappa; some uses, misuses, and alternatives. *Educational and Psychological Measurement* 1981; **41**:687–699.
8. Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics* 1975; **31**:651–659.
9. Green SB. A comparison of three indexes of agreement between observers: proportion of agreement, G-index, and kappa. *Educational and Psychological Measurement* 1981; **41**:1069–1072.
10. Kraemer HC, Bloch DA. Kappa coefficients in epidemiology: an appraisal of a reappaisal. *Journal of Clinical Epidemiology* 1988; **41**:959–968.

11. Landis JR, Koch GG. A review of statistical methods in the analysis of data arising from observer reliability studies (Part I). *Statistica Neerlandica* 1975; **29**:101–123.
12. Light RJ. Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological Bulletin* 1971; **76**:365–377.
13. Maclure M, Willett WC. Misinterpretation and misuse of the Kappa statistic. *American Journal of Epidemiology* 1987; **126**:161–169.
14. Shrout PE. Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research* 1998; **7**:301–317.
15. Thompson WD, Walter SD. A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology* 1988; **41**:949–958.
16. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 1993; **46**:423–429.
17. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* 1990; **43**:543–549.
18. Hoehler FK. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. *Journal of Clinical Epidemiology* 2000; **53**:499–503.
19. Lantz CA, Nebenzahl E. Behavior and interpretation of the k statistic: resolution of the two paradoxes. *Journal of Clinical Epidemiology* 1996; **49**:431–434.
20. May SM. Modelling observer agreement–an alternative to kappa. *Journal of Clinical Epidemiology* 1994; **47**:1315–1324.
21. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 1971; **76**:378–382.
22. Kraemer HC. Extensions of the kappa coefficient. *Biometrics* 1980; **36**:207–216.
23. Kupper LL. On assessing interrater agreement for multiple attribute responses. *Biometrics* 1989; **45**:957–967.
24. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–174.
25. Janes CL. Agreement measurement and the judgment process. *Journal of Nervous and Mental Disease* 1979; **167**:343–347.
26. Spitznagel EL, Helzer JE. A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry* 1985; **42**:725–728.
27. Cicchetti DV, Heavens RJ. A computer program for determining the significance of the difference betweeen pairs of independently derived values of kappa or weighted kappa. *Educational and Psychological Measurement* 1981; **41**:189–193.
28. Donner A, Eliasziw M. Sample size requirements for reliability studies. *Statistics in Medicine* 1987; **6**: 441–448.
29. Donner A, Eliasziw M, Klar N. Testing the homogeneity of kappa statistics. *Biometrics* 1996; **52**:176–183.
30. Donner A. Sample size requirements for the comparison of two or more coefficients of inter-observer agreement. *Statistics in Medicine* 1998; **17**:1157–1168.
31. Kraemer HC. Ramifications of a population model for k as a coefficient of reliability. *Psychometrika* 1979; **44**:461–472.
32. Aickin M. Maximum likelihood estimation of agreeement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics* 1990; **46**:293–302.
33. Maxwell AE. Coefficients of agreement between observers and their interpretations. *British Journal of Psychiatry* 1977; **130**:79–83.
34. Kraemer HC. Estimating false alarms and missed events from interobserver agreement: comment on Kaye. *Psychological Bulletin* 1982; **92**:749–754.
35. Cronbach LJ, Gleser GC, Nanda H, Rajaratnam J. *The Dependability of Behavioral Measurements*. Wiley: New York, 1972.
36. Lord FM, Novick MR. *Statistical Theories of Mental Test Scores*. Addison-Wesley: Reading, MA, 1968.
37. Kraemer HC. Measurement of reliability for categorical data in medical research. *Statistical Methods in Medical Research* 1992; **1**:183–199.
38. Carey G, Gottesman II. Reliability and validity in binary ratings. *Archives of General Psychiatry* 1978; **35**: 1454–1459.
39. Huynh H. Reliability of multiple classifications. *Psychometrika* 1978; **43**:317–325.
40. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 1968; **70**:213–229.
41. Darroch JN, McCloud PI. Category distinguishability and observer agreement. *Australian Journal of Statistics* 1986; **28**:371–388.
42. Donner A, Eliasziw MA. A hierarchical approach to inferences concerning interobserver agreement for multinomial data. *Statistics in Medicine* 1997; **16**:1097–1106.
43. Fleiss JL. *Statistical Methods For Rates and Proportions*. Wiley: New York, 1981.
44. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 1973; **33**:613–619.

45. Fleiss JL, Davies M. Jackknifing functions of multinomial frequencies, with an application to a measure of concordance. *American Journal of Epidemiology* 1982; **115**:841–845.
46. Donner A, Eliasziw M. A goodness-of-fit approach to inference procedures for the kappa statistics: confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine* 1992; **11**:1511–1519.
47. Bloch DA, Kraemer HC. 2×2 kappa coefficients: measures of agreement or association. *Biometrics* 1989; **45**:269–287.
48. Landis JR, Koch GG. A one-way components of variance model for categorical data. *Biometrics* 1977; **33**: 671–679.
49. Hanley JA. Standard error of the kappa statistic. *Psychological Bulletin* 1987; **102**:315–321.
50. Fisher RA. *Statistical Methods for Research Workers*, 2nd edn. Oliver & Boyd: London, 1928.
51. Efron B. Bootstrap methods: another look at the jackknife. *Annals of Statistics* 1979; **7**:1–26.
52. Flack VF. Confidence intervals for the interrater agreement measure kappa. *Communications in Statistics—Theory and Methods* 1987; **16**:953–968.
53. Fleiss JL, Cicchetti DV. Inference about weighted kappa in the non-null case. *Applied Psychological Measurement* 1978; **2**:113–117.
54. Blackman NJ-N, Koval JJ. Interval estimation for Cohen's kappa as a measure of agreement. *Statistics in Medicine* 2000; **19**:723–741.
55. Borenstein M. Hypothesis testing and effect size estimation in clinical trials. *Annals of Allergy*, *Asthma*, *and Immunology* 1997; **78**:5–16.
56. Kraemer HC. *Evaluating Medical Tests*: *Objective and Quantitative Guidelines*. Sage Publications: Newbury Park, CA, 1992.
57. Collis GM. Kappa, measures of marginal symmetry and intraclass correlations. *Educational and Psychological Measurement* 1985; **45**:55–62.
58. Kraemer HC, Kazdin AE, Offord DR, Kessler RC, Jensen PS, Kupfer DJ. Measuring the potency of a risk factor for clinical or policy significance. *Psychological Methods* 1999; **4**:257–271.
59. Ross DC. Testing patterned hypotheses in multi-way contingency tables using weighted kappa and weighted chi square. *Educational and Psychological Measurement* 1977; **37**:291–307.
60. Cohen J. The cost of dichotomization. *Applied Psychological Measurement* 1983; **7**:249–253.
61. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Wadsworth & Brooks/Cole Advanced Books & Software: Monterey, CA, 1984.
62. Noda AM, Kraemer HC, Yesavage JA, Periyakoil VS. How many raters are needed for a reliable diagnosis? *International Journal of Methods in Psychiatric Research* 2001; **10**:119–125.
63. Kraemer HC. How many raters? Toward the most reliable diagnostic consensus. *Statistics in Medicine* 1992; **11**:317–331.
64. Kraemer HC. What is the 'right' statistical measure of twin concordance (or diagnostic reliability and validity)? *Archives of General Psychiatry* 1997; **54**:1121–1124.
65. Kraemer HC. Statistical issues in assessing comorbidity. *Statistics in Medicine* 1995; **14**:721–733.
66. Tanner MA, Young MA. Modeling agreement among raters. *Journal of the American Statistical Association* 1985; **80**:175–180.