J Clin Epidemiol Vol. 49, No. 7, pp. 775–782, 1996

ELSEVIER

0895-4356/96/$15.00
PII S 0895-4356(96)00011-X

The Meaning of Kappa: Probabilistic Concepts of Reliability and Validity Revisited

Irene Guggenmoos-Holzmann

INSTITUTE OF MEDICAL STATISTICS AND INFORMATION SCIENCE,
FREIE UNIVERSITÄT BERLIN, D-12 200 BERLIN, GERMANY

ABSTRACT. A framework—the "agreement concept"—is developed to study the use of Cohen's kappa as well as alternative measures of chance-corrected agreement in a unified manner. Focusing on intrarater consistency it is demonstrated that for 2 × 2 tables an adequate choice between different measures of chance-corrected agreement can be made only if the characteristics of the observational setting are taken into account. In particular, a naive use of Cohen's kappa may lead to strinkingly overoptimistic estimates of chance-corrected agreement. Such bias can be overcome by more elaborate study designs that allow for an unrestricted estimation of the probabilities at issue. When Cohen's kappa is appropriately applied as a measure of chance-corrected agreement, its values prove to be a linear—and not a parabolic—function of true prevalence. It is further shown how the validity of ratings is influenced by lack of consistency. Depending on the design of a validity study, this may lead, on purely formal grounds, to prevalence-dependent estimates of sensitivity and specificity. Proposed formulas for "chance-corrected" validity indexes fail to adjust for this phenomenon. J CLIN EPIDEMIOL 49;7:775–782, 1996.

KEY WORDS. Diagnostic test, reliability, validity, kappa, chance-corrected agreement, chance-corrected validity

INTRODUCTION

Cohen's kappa [1] is a measure of reliability that has been used abundantly in a variety of different settings [2]. Because it is defined predominantly by a formal calculus, it remains difficult to judge whether this index conveys any meaningful information on what it is supposed to measure [3,4]. Literature on obviously contraintuitive results is immense [5–9]. This pertains to the dependence of kappa on prevalence as well as to the paradoxical behavior when calculation recurs on asymmetric tables. The difficulties in solving the observed paradoxes and the lack of a pragmatic concept of reliability behind kappa becomes even more striking when the wide range of possible definitions is taken into account that, like Cohen's kappa, quantify reliability in terms of "agreement beyond chance" [4, 10,11].

The mathematical background for the assessment of errors in qualitative classifications is usually taken from the theory of quantitative measurements [4]. Despite its fruitfulness in detail, it has stressed more the analogies than the differences between qualitative and quantitative measurements. A new melody has come into this theoretical framework by latent class analysis, which not only adds an arsenal of sophisticated statistical methodology but, more importantly, encourages new views through old perspectives [11–14].

In the following, a concept is developed in which the properties of different formulations for indexes of chance-corrected agreement can be studied in a unified way. The notion of chance-corrected agreement is shown to correspond to the assumption that inconsistencies in ratings are due to the fact that raters sometimes are indecisive on how to classify items. This interpretation of inconsistency has occasionally occurred in the literature: in a paper on errors in diagnosing dental caries, Lu [15] suggested a way to distinguish between judgments based on clear criteria and judgments that are pure guesswork because the application of criteria fails. Lu assumed that guessing is characterized by classifying teeth as carious with a probability of 1/2. It was noted by Maxwell [16] that Cohen's kappa implies the probability of positive guesses to be equal to the overall probability of positive ratings. Aickin [11] used a differentiation of items into those that are easy and those that are difficult to classify when he derived a special kappa-like measure of agreement. None of these authors has further pursued the more general implications of indecisiveness in the case of unclear items. From the viewpoint of latent class theory the distinction between clearcut and uncertain classifications of items is just one of many possible latent class models. Interestingly, this concept has been studied casually when dealing with the assessment of the validity of multiple ratings [17]. Although its characteristics have never been investigated in the context of reliability it can be shown to have a special bearing because it fosters a deeper understanding of the meaning of chance in kappa-like indexes of chance-corrected agreement. This is made explicit by focusing on the reliability of replicated dichotomous ratings.

EXAMPLE 1

Suppose a clinician performs 2 independent examinations of 100 ultrasound scans to assess the consistency of ratings in terms of the absence or presence of a specific lesion. The results are as shown in Table 1. The proportion of observed agreement is $P_o = (13 + 75)/100 = 0.88$. To judge this against the amount of agreement that would have occurred by chance, Table 1 is contrasted with a table of random agreement (Table 2). Customarily, this table is assumed

Address reprint requests to: Prof. Dr. Dr. Irene Guggenmoos-Holzmann, Institut für Medizinische Statistik und Informationsverarbeitung, Universitätsklinikum Benjamin Franklin, Freie Universität Berlin, Hindenbrugdamm 30, D-12 200 Berlin, Germany.

Accepted for publication on 29 August 1995.

TABLE 1 Agreement between two interpretations of the same 100 ultrasound scans by one clinician

		Second rating +	Second rating −	
First rating	+	13	7	20
	−	5	75	80
		18	82	100

to have the same marginal counts (Table 2a) as the observed agreement table. The proportion of random agreement in this table is $P_e = (3.6 + 65.6)/100 = 0.69$.

The amount of chance-corrected agreement, then, is calculated as Cohen's kappa:

$$\kappa_{Cohen} = (P_o - P_e)/(1 - P_e). \quad (1)$$

When applied to the example, $\kappa_{Cohen} = (0.88 - 0.69)/(1 - 0.69) = 0.61$. The expected proportion of randomly consistent ratings within Table 1 is $(P_o - \kappa_{Cohen}) = 0.88 - 0.61 = 0.27$. Obviously, only a part of Table 2a is used to explain random agreement within Table 1. This part is given by $(1 - \kappa_{Cohen})$, as may be seen from the following reformulation of Cohen's formula:

$$P_o - \kappa_{Cohen} = (1 - \kappa_{Cohen}) * P_e. \quad (2)$$

Thus, we must multiply the table of random agreement (Table 2a) by $(1 - \kappa_{Cohen}) = 1 - 0.61 = 0.39$ when we want to know the expected numbers of randomly consistent and inconsistent ratings (Table 2b) within the observed agreement table.

Even if one accepts that randomly combined ratings constitute a part of the cross-classification, one may question the proportion of positive ratings occurring by chance. One could argue that the *a priori* probability of positive ratings would have been 1/3. The table of randomly combined ratings, then, is as in Table 3a. The proportion P_e of chance agreement is $(1/9) + (4/9) = 0.56$. Using Cohen's formula for this P_e, we get $(0.88 - 0.56)/(1 - 0.56) = 0.73$ as the amount of chance-corrected agreement in the observed table. Using the same argument as above, Table 3a must be multiplied by $1 - 0.73 = 0.27$ to arrive at the cross-classification of random ratings expected to be contained in the observed cross-classification (Table 1).

Table 3b as well as Table 2b support the notion that discordant

TABLE 2. Chance agreement when the proportion of positive ratings is the same as in Table 1

a			b. Multiplied by 0.39	
	+	−	+	−
+	3.6	16.4	1.4	6.4
−	14.4	65.6	5.6	25.6
	100		39	

TABLE 3. Chance agreement when the proportion of positive ratings is 1/3

a			b. Multiplied by 0.27		
	+	−		+	−
+	11.1	22.2	+	3	6
−	22.2	44.4	−	6	12
	100			27	

ratings in Table 1 are due to random disagreement. Comparing the two resulting estimates of chance-corrected agreement, 0.61 and 0.73, the latter has the disadvantage of being based on a rather arbitrary *a priori* assumption. But is the assumption underlying the "correct" kappa value less arbitrary?

THE AGREEMENT CONCEPT

In the following, we use probabilities instead of counts or proportions to simplify formulas and facilitate argumentation. As in the example above, the focus will be on replicated classifications. In this case the marginals of the agreement table can be assumed to be balanced. The probabilistic notation of the resulting cross-classification of repeated ratings is given in Table 4.

$P_o = A + D$ is the probability of observed agreement. As assumed in Cohen's original formula, the probability of random agreement is defined by $P_e = Q^2 + (1 - Q)^2$. This refers to the marginal probability Q of the agreement table and characterizes random agreement as produced by ratings that (1) are positive with probability Q and (2) combine randomly when repeated on the same subject. While condition (2) directly corresponds to the concept of random agreement, the plausibility of condition (1) is less obvious. The question is if the choice of Q is a necessary ingredient of the procedure.

When we assume that κ is nonnegative, a reformulation of formula (1)

$$P_o = \kappa(1 - P_e) + P_e = \kappa + (1 - \kappa)P_e \quad (3)$$

shows that Cohen's concept of chance-corrected agreement can be interpreted as the partition of observed agreement P_o into random agreement P_e and systematic agreement where the probability of systematic agreement is κ. It would be contradictory to assume that the systematic agreement of the rater on an item was a random effect.

Therefore, formula (3) suggests that observations can be separated

TABLE 4. Agreement table

		Test 2 +	Test 2 −	
Test 1	+	A	B	Q
	−	C	D	1 − Q
		Q	1 − Q	1

TABLE 5. Decomposition of Table 1 according to two latent classes: "Systematic agreement" and "random agreement"

Systematic agreement				Random agreement		
κ^*	v		$+ (1-\kappa)^*$	w^2	$w(1-w)$
		$(1-v)$		$w(1-w)$	$(1-w)^2$

into a part V on which the rater systematically agrees and a part I of some sort of inconclusive observations on which the rater agrees only randomly. The former can be further divided according to whether the rating is systematically positive (V_{pos}) or systematically negative (V_{neg}). The results of ratings on items of part I are either randomly positive or negative, and the probability of positive ratings on these items can be quantified by p. This concept of chance-corrected agreement will be referred to here as the "agreement concept."

FORMAL APPROACH TO THE DEFINITION OF DIFFERENT INDEXES OF CHANCE-CORRECTED AGREEMENT

According to the agreement concept underlying formula (3), the resulting agreement table (Table 4) can be split into two tables (Table 5): in the first table each pair of ratings yields the same result that is positive with probability v. The second table is an independent cross-classification of ratings that are positive with probability w.

Suppose we knew that the probabilities $P(V)$ and $P(V_{pos})$ of systematic and systematically positive ratings equal 0.7 and 0.2, respectively, and we were free to choose the assignment probability p of inconclusive observations. For example, let p equal the proportion of systematically positive ratings within all systematic ratings: $p = 0.2/0.7$, or throw a coin so that $p = 1/2$, or use a die and rate positive if six dots come up. For each of these strategies, Table 5 and the correspondence between the $P(V)$, $P(V_{pos})$, p and κ, v, w can be used to arrive at a specific agreement table (Table 6). The differences between these tables are not too striking, yet visible. For all these tables, Cohen's kappa has been calculated. Only for the table belonging to assignment probability $p = 0.2/0.7$ does Cohen's kappa give an unbiased estimate of the probability $P(V)$ of systematic agreement. Note that in this case p equals the marginal probability Q of the agreement table. For the other tables Cohen's kappa either underestimates or overestimates the true probability $P(V)$ of chance-corrected agreement.

TABLE 6. Intraobserver agreement for different probabilities of randomly positive rating when the proportion of systematic ratings is 70% and the proportion of systematically positive ratings is 20%

$p = 0.2/0.7$		$p = 1/2$		$p = 1/6$	
0.225	0.061	0.275	0.075	0.208	0.042
0.061	0.652	0.075	0.575	0.042	0.708
$\kappa_{Cohen} = 0.70$		$\kappa_{Cohen} = 0.67$		$\kappa_{Cohen} = 0.78$	

Indeed, it depends crucially on the assumptions made on p whether Cohen's kappa or any other kappa-like measure yields an unbiased estimate of $P(V)$. To calculate κ from the observed Table 4, we have the two observed agreement cells with probabilities A and D that, according to Table 5, can be decomposed into

$$A = \kappa v + (1 - \kappa)w^2 \tag{4}$$

$$D = \kappa(1 - v) + (1 - \kappa)(1 - w)^2. \tag{5}$$

Because the other two cells of Table 4 are directly related to A and D by the equation $B = C = (1 - A - D)/2$, we have only two equations to solve for three unknown probabilities κ, v, and w. Therefore, a solution of κ will be found only if any assumptions are made on v or w. Indeed, if the association in Table 4 is positive, there is a clear relationship between a specific κ and the imposed restrictions on v or w. On the basis of this relationship, an infinite number of kappa-like indexes could be constructed. Three well-known examples are

$$\kappa_{Cohen} = \frac{P_o - [Q^2 + (1 - Q)^2]}{1 - [Q^2 + (1 - Q)^2]} \quad \text{if} \quad w = v \tag{6}$$

$$\kappa_{Aickin} = P_o\left(1 - \frac{1}{\sqrt{OR}}\right) \quad \text{if} \quad v = \frac{w^2}{1 - 2w + 2w^2} \tag{7}$$

$$\kappa_{0.5} = 2P_o - 1 \quad \text{if} \quad w = 0.5. \tag{8}$$

Note that for Cohen's kappa the equality $w = v$ implies that w equals the marginal probability Q of Table 4. Aickin [11] developed a version of a kappa-like measure directly on the basis of the decomposition in Table 5 by assuming that the odds $v/(1 - v)$ of systematic agreement are the same as the odds $w^2/(1 - w)^2$ of random agreement. Aickin's maximum likelihood solution of κ can be written in the closed form given for κ_{Aickin}, with odds ratio $OR = AD/BC$ [18]. The kappa index $\kappa_{0.5}$ for an assignment probability of 1/2 has first been proposed by Holley and Guilford [10]. This formulation of kappa has provoked criticism because it "is merely a linear transformation of P_o and so incorporates no adjustment for change" [19]. Yet, within the framework of the agreement concept it is evidently just one example of a specified assignment probability that is independent of the probability of systematically positive ratings. The marginal probability Q of the resulting agreement table thus can differ from 0.5. As exemplified in Table 6, any other fixed assignment probability w different from 1/2 may be used as well.

EXAMPLE 2

It was demonstrated above that a chosen index of chance-corrected agreement is unbiased if the probability of assigning inconclusive observations corresponds to the restrictions imposed by this index. The need for such restrictions is specific for 2×2 tables only. However, the probabilities can be estimated when, for example, more than two rating categories are chosen or more than two repetitions of each rating are performed. Then the probabilities involved in the agreement concept can be estimated by techniques of latent class analysis.

A set of 70 specimens of transsectal ultrasound-guided prostate biopsies was examined by a pathologist three times in random order to assess the presence of cancer. The pathologist rated the specimens during the hours of routine work and no more than one specimen

TABLE 7. Numbers of positive findings in three repeated assessments of 70 prostate biopsies

Number *i* of positive findings	*n*
2	19
1	2
1	9
0	30
Total:	70

was assessed per day. The results are displayed in Table 7. The version of Cohen's kappa for multiple ratings [3] yields a value of 0.79 for Table 7.

Again, using probabilities instead of counts the decomposition in a table of systematic ratings and a table of random ratings is given by the formula

$$Y_i = \kappa v_i + (1 - \kappa)\frac{3!}{i!(3-i)!}w^i(1-w)^{(3-i)} \qquad (9)$$

where Y_i is the probability of i positive findings in three replicated observations, w is the assignment probability of unclear items, and $v_1 = v_2 = 0$ so that v_3 and $v_0 = 1 - v_3$ are the probabilities of systematically positive and negative ratings. Because $Y_0 = 1 - Y_3 - Y_2 - Y_1$, there are three unknown probabilities, κ, $v = v_3$, and w, to be estimated from three equations. The solutions could be computed directly. A maximum likelihood estimation of the model parameters was performed with the aid of the GLIM-macro of Ekholm *et al.* [20] to enable a comparison between the estimates of this full model and models with restrictions similar to those considered in the case of 2 × 2 tables:

Variable	Full model	Restricted model (Cohen)	Restricted model (Aickin)	Restricted model (1/2)
κ	**0.65 (0.20)**	0.790 (0.06)	0.791 (0.07)	0.792 (0.07)
v	**0.64 (0.21)**	0.45 (0.05)	$= \frac{w^2}{1 - 2w + 2w^2}$	0.49 (0.09)
w	**0.18 (0.12)**	$= v$	0.48 (0.04)	$= 1/2$
Deviance (df)	**0.0 (0 df)**	4.06 (1 df)	11.63 (1 df)	16.68 (1 df)

For the full model, the probability κ of chance-corrected agreement is estimated as 0.65. The estimates of w and v differ by 46% although their standard errors are large. When it is assumed that the probability of systematically positive ratings equals that of randomly positive ratings (Cohen model), the estimate of κ becomes considerably larger than in the full model. Note that the estimate of κ given by this restricted model equals the multirater version of Cohen's kappa. Yet, Cohen's model exhibits a significant lack of fit to the data ($\chi^2 = 4.06$ with 1 df). Thus, this is an example in which the assumption underlying Cohen's kappa is clearly violated. Estimates of chance-corrected agreement are almost identical to that of Cohen's model when the restriction of Aickin's index or the restriction $v = 1/2$ is applied. Yet the fit of these models is even worse than the fit of Cohen's model. If in this setting the ratings had been repeated only twice it would have been overlooked that each of the common indexes overestimates the amount of chance-corrected agreement.

DEPENDENCE OF CHANCE-CORRECTED AGREEMENT ON PREVALENCE

The necessity of choosing the correct index of chance-corrected agreement in the case of 2 × 2 tables may become even clearer when the dependence of kappa-like indexes on prevalence is investigated. Let us consider two distinct populations Θ_0 and Θ_1 characterized by the respective absence and presence of a feature that the rating tries to grasp. Again, each of these populations may be partitioned into groups V^0_{pos}, V^0_{neg}, I^0 and V^1_{pos}, V^1_{neg}, I^1 of subjects that are systematically or randomly classified by the rater. The probability of systematically positive and negative ratings will be different in Θ_0 and Θ_1, and this may also be the case for inconclusive items. Usually, the items will be sampled from a mixture of these two populations. Thus, when in the sample the true prevalence of the feature under study is θ, and the parts of systematic and random ratings are denoted by V^θ_{pos}, V^θ_{neg}, I^θ, the probability of systematically positive ratings is $P(V^\theta_{pos}) = \theta P(V^1_{pos}) + (1 - \theta)P(V^0_{pos})$. This applies similarly for random ratings $P(I^\theta) = \theta P(I^1) + (1 - \theta)P(I^0)$ and for the whole set of systematic ratings $P(V^\theta) = \theta P(V^1) + (1 - \theta)P(V^0)$. That is, $P(V^\theta)$ is a linear function of prevalence. Therefore, if a specified κ is an unbiased estimate of $P(V^\theta)$, κ itself must be linear in θ. This proposed linear dependence of κ on prevalence is in striking contradiction to the parabolic shape of the curve that is usually displayed as having a maximum at a medium prevalence and approaching minimal values or even zero when prevalence is near 0 or 1. In the framework of the agreement concept, such a nonlinear dependence on prevalence is explicable only by a conflict between an underlying assignment probability p and the κ formula chosen to estimate $P(V^\theta)$.

Because, in reality, nothing is known about the true state of the items, the assignment of inconclusive items I_θ can at the most be guided by the probability $P(V^\theta_{pos})$ of systematically positive ratings. For example, associated with Cohen's concept of "chance" is the idea that p exactly equals the proportion of systematically positive ratings: $p = P(V^\theta_{pos})/P(V^\theta)$. Similarly, the assignment probability corresponding to κ_{Aickin} is a function of $P(V^\theta_{pos})/P(V^\theta)$. In contrast, the probability corresponding to $\kappa_{0.5}$ is clearly independent of the relative amount of systematically positive ratings. Thus, although the assignment of inconclusive items is not directly influenced by the prevalence of the true state, it is related to this prevalence through its dependence on $P(V^\theta_{pos})$ and $P(V^\theta)$. What happens if a chosen measure κ of chance-corrected agreement does not correspond to the underlying probability of assigning inconclusive items?

In Fig. 1, the straight line G represents the unbiased estimates that are given by κ_{Cohen} when the underlying assignment probability p equals $P(V^\theta_{pos})/P(V^\theta)$. The same straight line results for any other kappa index provided that the assignment probability generating the prevalence-dependent agreement table corresponds to that kappa index. In particular, it represents the curve of $\kappa_{0.1}$, $\kappa_{0.5}$, and $\kappa_{0.9}$ when the respective assignment probabilities p are 0.1, 0.5, and 0.9. The other curves show the values of Cohen's kappa that would result if the data were generated by just these three fixed assignment probabilities. If the assignment probability p is 1/2, Cohen's kappa underestimates the true probability $P(V^\theta)$. With more extreme assignment probabilities Cohen's kappa increasingly overestimates the amount of agreement beyond chance.

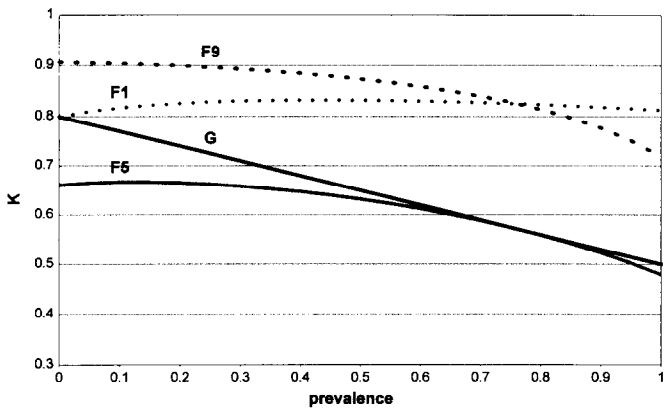


FIGURE 1. Cohen's kappa as a function of the true prevalence and the probability p of randomly positive rating.

$P(V^0) = 0.8$, $P(V^1) = 0.5$, $P(V^0_{pos}) = 0.08$, $P(V^1_{pos}) = 0.35$

G, unbiased estimates of kappa indexes that correctly correspond to p: ($\kappa_{0.1}$, $p_{0.1}$), ($\kappa_{0.5}$, $p_{0.5}$), ($\kappa_{0.9}$, $p_{0.9}$), (κ_{Cohen}, p_{Cohen}); F1, estimates of κ_{Cohen} if $p = 0.1$; F5, estimates of κ_{Cohen} if $p = 0.5$; F9, estimates of κ_{Cohen} if $p = 0.9$.

To produce a parabolic curve for the dependence of Cohen's kappa on prevalence it must be assumed that the assignment probabilities for items in I^0 and I^1 are different. This may happen if there are two types of inconclusive features that are associated with the true state and evoke different strategies of assignment. Then, Cohen's kappa will overestimate the true amount of systematic agreement over the whole spectrum of prevalences (Fig. 2).

Even for an inappropriately chosen index of chance-corrected agreement there is no formal necessity for small values in populations with extreme prevalence. This supports the observation of Shrout *et al.* [21] that large values of Cohen's kappa can be found even in populations with very small trait prevalence.

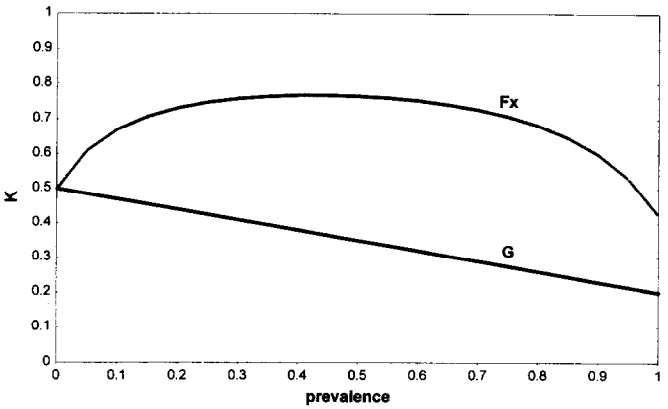


FIGURE 2. Cohen's kappa as a function of the true prevalence and the probability p of randomly positive rating.

$P(V^0) = 0.5$, $P(V^1) = 0.2$, $P(V^0_{pos}) = 0.05$, $P(V^1_{pos}) = 0.13$

G, unbiased estimates; Fx, κ_{Cohen} if $p = 0.1$ for items in I^0 and $p = 0.9$ for items in I^1.

THE AGREEMENT CONCEPT AND THE CLASSIC CONCEPT OF RELIABILITY

The special features of the agreement concept become clear when compared with classic measurement theory. Here, in the most abstract formulation, the reliability of repeated measurements on a set of items is defined as the variance of the average measurements per item divided by the variance of all measurements. The classic intraclass correlation coefficient ICC is a measure of reliability. For normally distributed measurements, it allows an interpretation of a single measurement Y_{ij} of rater i on subject j as arising by an erroneous deviation ε_{ij} from an ideal measurement X_i provided that errors are independent of X_i and normally distributed with mean 0. Within this framework, reliability depends on the variability of the ideal measurements X_i and the amount of measurement error.

By interpreting dichotomous ratings as quantitative measurements with values 0 and 1, the ideal measurement X_i is viewed as a consensus score s_i that arises as the mean of repeated binary ratings per subject or, equivalently, as the probability of a positive rating when ratings are repeated on subject i. The intraclass correlation coefficient, then, can be written as

$$\text{ICC} = \frac{\text{var}(s_i)}{Q(1 - Q)} \tag{10}$$

where Q is, as before, the marginal probability of positive findings in the cross-classification of repeated ratings [8]. This formulation of the intraclass correlation coefficient for two replicate ratings is equivalent to Cohen's kappa [3]. To avoid ambiguity the abbreviation ICC is used for Cohen's kappa as a measure of classic reliability.

In the agreement concept, measurement error is assumed to occur due to inconclusive observations. The relation between the intraclass correlation coefficient ICC and the probability $P(I)$ of these inconclusive observations is given by

$$\text{ICC} = \frac{Q(1 - Q) - P(I)p(1 - p)}{Q(1 - Q)} = 1 - P(I)\frac{p(1 - p)}{Q(1 - Q)} \tag{11}$$

where p is the probability of rating an inconclusive item as positive. It has been shown above that Cohen's kappa is an unbiased estimate of chance-corrected agreement only if $p = Q$. Formula (11), then, reduces to $\text{ICC} = 1 - P(I) = P(V)$. Thus, only in this case is classic reliability, as measured by the intraclass correlation coefficient, identical to the measure of chance-corrected agreement.

Moreover, formula (11) shows that Cohen's kappa as a measure of classic reliability is a mixture of two components: the amount of certainty encountered in the rating process and the probability of assigning an unclear item to one of the two rating categories. A consequence is that reliability can be increased in two ways, either by clarifying the inconclusiveness of items or by changing the assignment probability. Figure 3 shows that any degree of reliability in the classic sense can be achieved by choosing an extreme assignment probability while the amount of inconclusive observations remains unchanged. When judging the quality of improvements in (classic) reliability it may be of practical importance to differentiate between improvements in the amount of systematic ratings and mere changes in the guessing strategy.

ON THE VALIDITY OF UNRELIABLE RATINGS

It is common knowledge that the validity of diagnostic classifications is prone to be impaired by a lack of reliability. Yet, when valid-

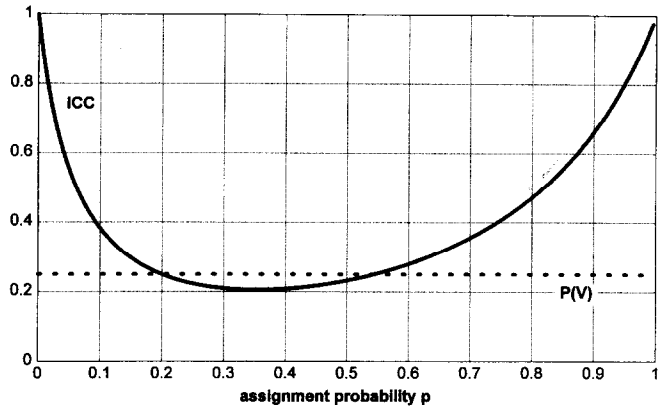


FIGURE 3. Dependence of Cohen's kappa as an intraclass correlation coefficient (ICC) on the probability of random assignment when $P(V) = 0.25$ and $P(V_{pos}) = 0.2$.

ity indexes such as sensitivity or specificity are estimated on the basis of a clinical investigation, the problem of reliability is rarely at issue. The preponderance of validity is substained by the impression that lack of reliability is responsible only for a relatively small part of overall misclassification. Indeed, this impression may be biased. In any case, it is not supported by the probabilistic models on which the assessment of validity and reliability commonly is based.

Let us assume that the pathological feature that an observer aims to classify as present or absent is clearly and uniquely defined. If the ratings were absolutely reliable, then meticulously planned and performed validity studies in populations with different true prevalence all would yield the same sensitivity SE* and the same specificity SP* of the rating. What is the impact of chance agreement on these validity estimates? When chance agreement is measured by a kappa-like index, then it may be assumed that the corresponding probabilistic conceptualization applies.

According to the agreement concept the error-inflated sensitivity and specificity can be expressed in terms of systematic and random rating:

$$SE = P(V^1_{pos}) + P(I^1)p \quad (12)$$

$$SP = P(V^0_{neg}) + P(I^0)(1 - p). \quad (13)$$

Within the agreement concept, the natural definition of chance-corrected validity consists of the probability of correct ratings that are not due to random assignment:

$$SE^* = P(V^1_{pos}) = SE - P(I^1)p$$

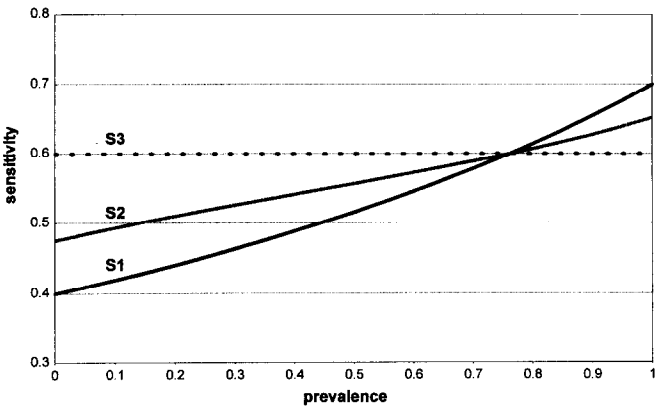
$$SP^* = P(V^0_{neg}) = SP - P(I^0)(1 - p).$$

Because the probabilities $P(I^1) = 1 - P(V^1)$ and $P(I^0) = 1 - P(V^0)$ can only be estimated by some kappa-like index in a concomitant study of intraobserver agreement, a validity study without such additional assessment of reliability will not arrive at any meaningful measure of chance-corrected validity.

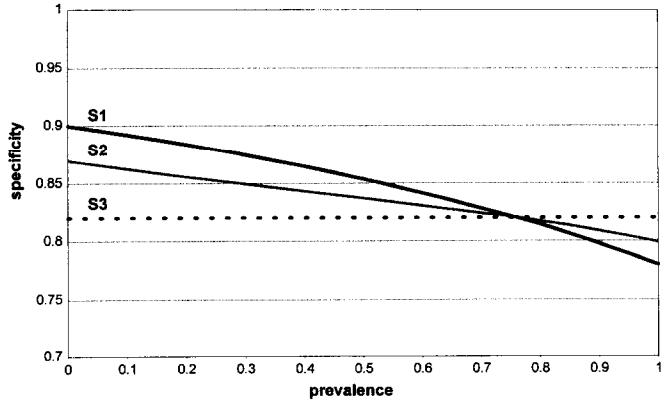
Whereas for absolutely reliable ratings indexes of validity ideally are independent of prevalence, such independence is questionable when the inconsistency of ratings is taken into consideration. As has been shown above, Cohen's kappa as well as Aickin's measure of chance-corrected agreement are based on the assumption that the assignment probability p of inconclusive items depends on the prevalence of the feature under study. Because p is involved in formulas (12) and (13) as a characteristic of validity, it has a formal impact on observed sensitivity and specificity. For example, when in a given setting, Cohen's kappa is an unbiased measure of chance-corrected agreement, then the corresponding assignment probability p_{Cohen} will have the effect that sensitivity monotonously increases (Fig. 4a) and specificity monotonously decreases (Fig. 4b) with increasing prevalence. The same is true when p_{Aickin} applies in a given setting. Note that the decision rule underlying Cohen's kappa yields the highest maxima for specificity and sensitivity but also the lowest minima. The range of sensitivities is 40 to 70%, of specificities 78 to 90%, depending on prevalence. The validity of unreliable ratings is independent of prevalence if and only if the assignment probability p of inconclusive items is independent of prevalence, for example $p = 1/2$ (Fig. 4a and b).

DISCUSSION

When the question is whether an index of chance-corrected agreement sensibly gauges the consistency of ratings, it is not the formula of the index that yields the answer, but the concept underlying the formula. Only the concept can be explored with respect to its plausibility in practice. An important feature of the concept underlying



A



B

FIGURE 4. Sensitivity (a) and specificity (b) as functions of the true prevalence and the probability p of randomly positive rating.

$P(V^0) = 0.8$, $P(V^1) = 0.5$, $P(V^0_{pos}) = 0.08$, $P(V^1_{pos}) = 0.35$

S1, $p = p_{Cohen}$; S2, $p = p_{Aickin}$; S3, $p = 0.5$.

kappa-like indexes is the unspoken assumption that the act of categorizing, characterizing, or classifying is contaminated by an unconscious inclination to adhere to a classification even if an assignment to given categories is difficult. This assumption seems to be more difficult to accept in psychological [22] than in medical decision making (e.g., see Refs. 23 and 24), where the inclination toward guessing may be alleviated but not ruled out completely by the introduction of an additional category "intermediate, indeterminate, indefinite" [25,26].

In weighing the actual characteristics of a rating process, the dichotomy of systematic and random ratings proposed by the agreement concept must be compared with alternative concepts. One of these is characterized by the assumption that there are no unclear items but only erroneous deviations "from a 'true' assessment which [the rater] would make were he able to give limitless time and concentration to the task" [13]. Like the agreement concept, it is handled analytically as a latent class model. The latent classes are the "true" assignments, and the deviations from these assignments are estimated as error probabilities. Moreover, like the agreement concept, it allows for a more detailed analysis of reliability than the summary measure of reliability that is derived from classic measurement theory. It is up to philosophical or psychological debate on medical decision making as to which of the different concepts of a rating process is more realistic in a given context.

Cohen's kappa is used extensively for all kinds of qualitative data: ratings of pathologists or radiologists, clinicians' classifications of patient characteristics, answers to self-administered questionnaires, etc. Provided that the use of an index of chance-corrected agreement is conceptually justified, it should be accompanied by some thought on whether or not the assignment probability of random ratings is influenced by the proportion of systematically positive ratings. For example, it makes sense to take such influence into consideration if a rating process such as the interpretation of an electrocardiogram is part of the clinical routine. On the other hand, a correlation between positive ratings and assignment probability would be quite implausible if the reliability of a self-administered questionnaire is to be assessed. Because the respondents are barely able to have any impression on the proportion of positive answers in the rest of the sample, they are supposed to use a fixed assignment probability. If this assignment probability is correlated with the true responder status, Cohen's kappa will generally overestimate the chance-corrected agreement of responses (Fig. 2). In settings in which observers experimentally classify a sample of entities with predefined but unknown trait prevalence, the assignment probability of inconsistent items will usually be fixed, although the prior experience of the observer may influence the overall strategy of assignment. Quality assurance activities of pathologists or radiologists are examples of such settings. It should be kept in mind also that in experimental settings the estimate of chance-corrected agreement may be seriously biased if the type of assignment of inconclusive items does not correspond with the chosen kappa coefficient. Figure 1 and Example 2 demonstrate this phenomenon.

The risk of bias also pertains when a study aims at assessing the dependence of reliability on characteristics of the observed items [27]. Not only the choice of an adequate measure of chance-corrected agreement, but also the dependence of this measure on prevalence, are critical for the assessment of factor effects because a factor may "cause" differences in agreement just by the fact that different factor levels are associated with different trait prevalences.

The agreement concept has further implications on the interpretation of validity indexes emerging from more or less convenient samples. As was demonstrated above, a lack of reliability may on purely formal grounds lead to prevalence-dependent estimates of sensitivity and specificity. Several diagnostic procedures have been shown to vary in validity when applied to populations of different disease prevalences. This phenomenon has been ascribed to differences in disease characteristics and to a specific selection of cases in populations with differing prevalences. Furthermore, verification bias has been proposed as a cause of the observed varying validity [28–30]. Prevalence-dependent assignment probabilities of inconsistent observations may add to these effects of biased design. An example is provided by the investigations on exercise testing as a diagnostic marker for coronary artery disease [31]. Hlatky *et al.* [32] showed in a multifactorial analysis that the validity of the procedure was influenced by age, sex, type of chest pain, etc. Yet the target disease is known to be far more probable in older than in younger, and in male than in female, patients. Therefore, in an unblinded study, the increased sensitivity and decreased specificity in men and in older people are explainable not only by verification bias, but by an increased positive rating in basically inconclusive observations as well. A further finding of this analysis was that sensitivity appeared to be influenced by more factors than specificity. In terms of the agreement concept this phenomenon may simply be due to the fact that systematic ratings were less prevalent in patients with the disease than without the disease. As Fig. 4a and b show, this would result in a decreased prevalence-dependent variability of specificity and, thus, in a lower power when assessing factor effects on specificity.

The role of prevalence-dependent random rating can further be elucidated by comparing the validity estimates of blinded and unblinded studies [33]. Weintraub *et al.* [34] found that in a blinded study, a history of typical angina had no effect on validity while in an unblinded study, sensitivity was increased and specificity decreased when compared with ratings in patients with atypical chest pain. This conforms to the former consideration that assignment probabilities of inconclusive items can be taken as fixed and unaltered by prevalence in blinded studies, but should be suspected to be dependent on prevalence in unblinded studies with convenient clinical samples.

Work on "chance-corrected" validity indexes [27,35] seems to point to some solution of this problem. Yet the proposed formulas lack a conceptual background of what should be considered as "chance." It is the sparseness of information contained in a 2×2 table that prevents the modeling of reliability and validity in terms of random and systematic error. When a discrimination of the various sources of error is impossible, it is tempting to blend reliability and validity together. However, the interpretation of corresponding indexes is at best difficult. A more differentiated look at random and systematic error in qualitative data will enhance not only the planning but also the evaluation of diagnostic marker studies.

So far, the agreement concept has been discussed solely in the realm of intraobserver agreement. Most of the paradoxes in the application of Cohen's kappa, however, have arisen when assessing interrater consistency. Although the formula of Cohen's kappa is easily adapted to an asymmetric 2×2 agreement table, it is not a trivial exercise in the case of two raters to describe how the formula relates to the agreement concept. The concept singles out two features: the assignment probability p of inconclusive items and the set of items that is gauged by the probability of systematic agreement or "agreement beyond chance." This notion of systematic agreement refers, first of all, to a feature of intraobserver reliability: it describes the proportion of items that can clearly be categorized by

one observer. Thus, two observers can differ in several aspects: in the items they experience as easy to classify, and in the assignment probabilities they use in the case of uncertainty. Inconsistent ratings may not only be due to the fact that an item is inconclusive for both observers; inconsistencies may arise as well when only one of the observers is uncertain, or when both observers are certain but assign the item to different categories. To deal with these different types of systematic and random consistencies and inconsistencies, more information is needed than a 2×2 agreement table can provide. Kappa-like indexes that are based on these tables are prone to fail in assessing interobserver agreement correctly. Therefore, the necessity of more elaborate designs of interobserver studies must be acknowledged not only when other latent class approaches [13] but also when the agreement concept is used as a basis of argumentation.

This work was supported in part by BMFT Grant 07 PHF 01.

References

1. Cohen J. A coefficient of agreement for nominal scales. **Educ Psychol Meas** 1960; 20: 37–46.
2. Feinstein AR. A bibliography of publications on observer variability. **J Chron Dis** 1985; 38: 619–632.
3. Fleiss JL. Statistical Measures for Rates and Proportions, 2nd Ed. John Wiley & Sons, New York, 1981.
4. Kraemer HC. Ramifications of a population model for κ as a coefficient of reliability. **Psychometrika** 1979; 44: 461–472.
5. Brennan RL, Prediger DJ. Coefficient kappa: Some uses, misuses, and alternatives. **Educ Psychol Meas** 1981; 41: 687–699.
6. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. **Am J Epidemiol** 1987; 126: 161–169.
7. Thompson WD, Walter SD. A reappraisal of the kappa coefficient. **J Clin Epidemiol** 1988; 41: 949–958.
8. Kraemer HC, Bloch DA. Kappa coefficients in epidemiology: An appraisal of a reappraisal. **J Clin Epidemiol** 1988; 41: 959–968.
9. Feinstein AR, Cicchetti DV. High agreement but low kappa. **J Clin Epidemiol** 1990; 43: 543–549, 553–558.
10. Holley W, Guilford JP. A note on the G-index of agreement. **Educ Psychol Meas** 1964; 24: 749–753.
11. Aickin, M. Maximum likelihood estimation of agreement in the constant preditive probability model, and its relation to Cohen's kappa. **Biometrics** 1990; 46: 293–302.
12. Dawid AP, Skene AM. Maximum likelihood estimation of observer error-rates using the EM algorithm. **Appl Stat** 1979; 28: 20–28.
13. Baker SG, Freedman LS, Parmar MKB. Using replicate observations in observer agreement studies with binary assessments. **Biometrics** 1991; 47: 1327–1338.
14. Uebersax JS, Grove WM. Latent class analysis of diagnostic agreement. **Stat Med** 1990; 9: 559–572.
15. Lu KH. A critical evaluation of diagnostic errors, true increment and examiner's accuracy in caries experience assessment by a probabilistic model. **Archs Oral Biol** 1968; 13: 1133–1147.
16. Maxwell AE. Coefficients of agreement between observers and their interpretation. **Br J Psychiatry** 1977; 130: 79–83.
17. Formann AK. Measurement error in caries diagnosis: Some further latent class models. **Biometrics** 1994; 50: 865–875.
18. Guggenmoos-Holzmann I. How reliable are chance-corrected measures of agreement? **Stat Med** 1993; 12: 2191–2205.
19. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. **J Clin Epidemiol** 1993; 46: 423–429.
20. Ekholm A, Green M, Palmgren J. Fitting exponential family nonlinear models in GLIM 3.77. **GLIM Newlsett** 1986; 13: 4–13.
21. Shrout EP, Spitzer RL, Fleiss JL. Quantification of agreement in psychiatric diagnosis revisited. **Arch Gen Psychiatry** 1987; 44: 172–177.
22. Uebersax JS. Validity inferences from interobserver agreement. **Psychol Bull** 1988; 104: 405–416.
23. de Vet HCW, Koudstaal J, Kwee W, Willebrand D, Arends JW. Efforts to improve interobserver agreement in histopathological gradings. **J Clin Epidemiol** 1995; 48: 869–873.
24. Walter SD, Mitchell A, Southwell D. Use of certainty of opinion data to enhance clinical decision making. **J Clin Epidemiol** 1995; 48: 897–902.
25. Simel DL, Feussner JR, Delong ER, Matchar DB. Intermediate, indeterminate, and uninterpretable diagnostic test results. **Med Decision Making** 1987; 7: 107–114.
26. Feinstein AR. The inadequacy of binary models for the clinical reality of three-zone diagnostic decisions. **J Clin Epidemiol** 1990; 43: 109–113.
27. Coughlin SS, Pickle LW. Sensitivity and specificity-like measures of the validity of a diagnostic test that are corrected for chance agreement. **Epidemiology** 1992; 3: 178–181.
28. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. **N Engl J Med** 1978; 299: 926–930.
29. Feinstein AR. **Clinical Epidemiology:** The Architecture of Clinical Research. W. B. Saunders, Philadelphia, Pennsylvania, 1985.
30. Diamond GA. Reverend Bayes' silent majority: An alternative factor affecting sensitivity and specificity of exercise electrocardiography. **Am J Cardiol** 1986; 57: 1175–1179.
31. Philbrick JT, Horwitz RI, Feinstein AR. Methodologic problems of exercise testing for coronary artery disease: Groups, analysis and bias. **Am J Cardiol** 1980; 46: 807–812.
32. Hlatky MA, Mark DB, Harrell FE Jr, Lee KL, Califf RM, Pryor DB. Rethinking sensitivity and specificity. **Am J Cardiol** 1987; 59: 1195–1198.
33. Begg CB. Biases in the assessment of diagnostic tests. **Stat Med** 1987; 6: 411–423.
34. Weintraub WS, Madeira SW, Bodenheimer MM, Seelaus PA, Katz RI, Feldman MS, Agarwal JB, Banka VS, Helfant RH. Critical analysis of the application of Bayes' theorem to sequential testing in the noninvasive diagnosis of coronary artery disease. **Am J Cardiol** 1984; 54: 43–49.
35. Brenner H, Gefeller O. Chance-corrected measures of the validity of a binary diagnostic test. **J Clin Epidemiol** 1994; 47: 627–633.