Some common errors of experimental design, interpretation and inference in agreement studies

Erdmann, T. P.; De Mast, J.; Warrens, M. J.

Published in:
Statistical Methods in Medical Research

DOI:
10.1177/0962280211433597

Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2015, Vol. 24(6) 920–935

DOI: 10.1177/0962280211433597
smm.sagepub.com
SAGE

Some common errors of experimental design, interpretation and inference in agreement studies

TP Erdmann,[1] J De Mast[1] and MJ Warrens[2]

Abstract

We signal and discuss common methodological errors in agreement studies and the use of kappa indices, as found in publications in the medical and behavioural sciences. Our analysis is based on a proposed statistical model that is in line with the typical models employed in metrology and measurement theory. A first cluster of errors is related to nonrandom sampling, which results in a potentially substantial bias in the estimated agreement. Second, when class prevalences are strongly nonuniform, the use of the kappa index becomes precarious, as its large partial derivatives result in typically large standard errors of the estimates. In addition, the index reflects rather one-sidedly in such cases the consistency of the most prevalent class, or the class prevalences themselves. A final cluster of errors concerns interpretation pitfalls, which may lead to incorrect conclusions based on agreement studies. These interpretation issues are clarified on the basis of the proposed statistical modelling. The signalled errors are illustrated from actual studies published in prestigious journals. The analysis results in a number of guidelines and recommendations for agreement studies, including the recommendation to use alternatives to the kappa index in certain situations.

Keywords

agreement, association, inter-observer agreement, kappa, nominal measurement, reliability

1 Introduction

Diagnostic tests, clinical diagnoses and ratings can be perceived as measurements on a nominal scale. They classify subjects into a set of unordered categories, aiming to reflect an empirical property of the subjects that is not observed directly. This underlying property is often referred to as the 'true value' or 'actual state', but is called the 'measurand' in measurement theory and

[1]Institute for Business and Industrial Statistics of the University of Amsterdam, Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands.
[2]Institute of Psychology, Unit Methodology and Statistics, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands.

Corresponding author:
TP Erdmann, Institute for Business and Industrial Statistics of the University of Amsterdam, Plantage Muidergracht 12, 1018 TV Amsterdam, The Netherlands.
Email: t.p.erdmann@uva.nl

metrology (see, for example, the International Standardization Organization's *Guide to the Expression of Uncertainty in Measurement*[1]).

The quality or reliability of nominal measurements is often expressed in terms of agreement, typically in the form of a κ (kappa) index. Introduced by Cohen,[2] it is a measure of agreement between repeated classifications that corrects for agreement 'by chance', that is, for agreement achieved by blind classifications.[3–5] The κ index is surrounded by quite some controversy, and a number of papers have identified paradoxical behaviour.[6–12] Still, it has a prominent place in literature, education and practice in the social and medical sciences.

This article aims to signal and comment on a number of common methodological errors made in agreement studies and applications of the κ index found in scientific publications in prestigious journals. Some of these errors concern the design of agreement studies, and their ramifications include a potentially serious bias in the estimated κ index. Other errors are related to the standard error of the estimated agreement; the parameters of many agreement studies are such that this standard error, and consequently the confidence margins on the estimated κ index, is so large as to make the studies' results fairly useless. Finally, there are a number of interpretation pitfalls, stemming from the ambiguity of the concept of chance agreement, and from the strong under-weighting of the agreement on less prevalent classes in the case of strongly nonuniform class prevalences. As a consequence of these interpretation pitfalls, reported κ values may not reflect the authors' intention.

The next section gives a statistical model for nominal measurements, and defines the κ index in terms of this statistical model. Three clusters of problematic issues concerning the κ index are introduced, which are explained in the three subsequent sections and illustrated with examples of agreement studies in the literature of the social and medical sciences. The final section presents our recommendations for agreement studies. Throughout the article, issues are illustrated with actual papers from the literature. Please note that these examples are not in any way intended to discredit the work of their authors, but instead, are presented to draw attention to methodological shortcomings in practices in the medical and behavioural sciences.

2 Experimental design and statistical model

As we see it, many of the errors and much of the confusion to be discussed are rooted in the weak and superficial theoretical foundation of many expositions and discussions of agreement. Many discussions are framed in terms of sample statistics, without reference to a population model, and the few population models that are introduced do not capture what we see as essential characteristics of measurement and measurement reliability. We see the conceptualization and modelling that we present in this section as an important contribution to the theory of agreement studies.

To assess the quality of a nominal measurement procedure, one could collect data in the following manner. A sample of n subjects, randomly selected from the population of subjects, are independently classified once by each of m appraisers on an unordered scale $\{0, 1, \ldots, a-1\}$. The results are denoted Y_{ij}, with $i = 1, \ldots, n$ indexing subjects, and $j = 1, \ldots, m$ indexing appraisers. Measurements are intended to reflect an underlying empirical property of the subjects, named the measurand, and denoted X_i, which assumes the same values $\{0, 1, \ldots, a-1\}$. In the population of subjects, the measurand is assumed stochastically independent with a discrete distribution given by

$$p(k) = P(X_i = k),\ k = 0, \ldots, a-1, \text{ with } \sum_{k=0}^{a-1} p(k) = 1 \tag{1}$$

(the *class prevalences*). As for the distribution of the Y_{ij}, we assume that given a subject's true state X_i, the m measurements $Y_{i1}, Y_{i2}, \ldots, Y_{im}$ are stochastically independent (the assumption of *conditional independence*). Moreover, the distribution of the $Y_{i1}, Y_{i2}, \ldots, Y_{im}$ depends on X_i, and we define

$$q(k|l) = P(Y_{ij} = k|X_i = l), \tag{2}$$

thus specifying the distribution of the measurement errors. Note that, in case of a dichotomous test resulting in $Y=0$ (negative) or $Y=1$ (positive), $q(0|0)$ is the test's specificity (the probability of a correct negative), and $q(1|1)$ is the test's sensitivity (the probability of a correct positive), while $p(1)$ is the prevalence of the disorder. The model parameters $p(k)$, $k=0, 1, \ldots, a-1$, and $q(k|l)$, $k, l=0, 1, \ldots, a-1$, determine the distribution of the Y_{ij} and we have

$$P(Y_{ij} = k) = \sum_{l=0}^{a-1} p(l)q(k|l) = q(k) \text{ (marginal distribution).} \tag{3}$$

Situations may deviate from the assumptions above in numerous ways, and one's objectives may motivate alternative study designs. For example, the abovementioned assumption of conditional independence is often violated in practice due to nuisance factors affecting the results. For the purposes of our exposition, we think it is productive to keep the basic model relatively simple, and comment, where suitable, on possible extensions.

We now turn to the evaluation of nominal measurements in terms of a probability of agreement. Two measurements of a subject agree if they are identical (the subject is classified in the same category both times). $P_{Agreement}$ (or short: P_A) is the probability that two arbitrary measurements of an arbitrary subject agree. Under the model specified by Equations (1)–(2), we have for a subject with actual state $X_i=l$:

$$P_A(l) = P(Y_{ij_1} = Y_{ij_2}|X_i = l) = \sum_{k=0}^{a-1} q^2(k|l),$$

and for an arbitrary subject:

$$P_A = P(Y_{ij_1} = Y_{ij_2}) = \sum_{l=0}^{a-1}\sum_{k=0}^{a-1} p(l)q^2(k|l). \tag{4}$$

Fleiss[3] introduced the sample statistic

$$\hat{P}_A = \frac{1}{nm(m-1)}\sum_{i=1}^{n}\sum_{k=0}^{a-1} N_{ik}(N_{ik}-1)$$

where $N_{ik} = \{\#j\colon Y_{ij}=k\}$. De Mast and Van Wieringen[13] show that $\hat{P}_A$ is an unbiased estimator of P_A (that is, $\mathrm{E}\hat{P}_A = P_A$).

The probability of agreement is positive, even if measurements are completely unrelated to the measurand they intend to reflect. To correct for this phenomenon, Cohen,[2] Fleiss,[3] Conger[4] and numerous others have introduced κ indices as a recentred and rescaled version of P_A. They are defined as

$$\kappa = \frac{P_{\mathrm{A}} - P_{\mathrm{A|C}}}{1 - P_{\mathrm{A|C}}}, \tag{5}$$

where $P_{A|C}$ is the probability of agreement of random measurements 'by chance'. The value $\kappa=1$ corresponds to the agreement that a perfect measurement system would attain, and 0 corresponds to

the agreement that 'chance measurements' would attain. The most common conception of chance measurements is the one by Fleiss,[3] where chance measurements are defined as independent of the measurand with a probability distribution equal to the marginal distribution of the measurement system under study when applied to the subjects population under study (as given in (3)). Denoting chance measurements by Z_{ij}, this amounts to

$$Z_{ij} \text{ are i.i.d. and } P(Z_{ij} = k) = q(k) \text{ for all } i,\ j,\ \text{and } k.$$

Under these premises, the probability of agreement of chance measurements equals $P_{\mathrm{A|C}} = P(Z_{ij_1} = Z_{ij_2}) = \sum_{k=0}^{a-1} q^2(k)$.

Fleiss's[3] sample statistic

$$\hat{P}_{\mathrm{A|C}} = \sum_{k=0}^{a-1} \frac{N_k^2}{(mn)^2},$$

(with $N_k = \{\#(i, j)\colon Y_{ij} = k\}$) estimates $P_{\mathrm{A|C}}$ with a minor bias.[13] The sample $\hat{\kappa}$ is defined as in (5), but with the sample statistics $\hat{P}_A$ and $\hat{P}_{\mathrm{A|C}}$ instead of the corresponding population parameters.

Agreement studies are commonly done as part of scientific endeavour in the social and medical sciences (and beyond), and the results are frequently reported in the form of κ indices. Upon reviewing a large number of publications reporting on the results of agreement studies, we identified a number of commonly made methodological errors. We present and discuss these errors in the next sections, and we give examples of such errors in the existing literature. First, we speculate on the grounds for these errors.

The presentation and modelling above deviate from many expositions in the literature in two important aspects. First, they define an experimental model with population parameters, and define P_A and κ in terms of these population parameters. Sample statistics $\hat{P}_A$ and $\hat{\kappa}$ are presented as estimators for P_A and κ. In the literature, expositions are often framed in terms of sample statistics only, without referring to a population model (although there are some notable exceptions[11,14–16]). This makes it difficult to assess the properties of κ statistics as estimators of a population parameter. For example, inferences based on sample statistics should include an assessment of the estimate's standard error or confidence margins.

Another noteworthy characteristic of the given exposition, is that it attributes total dispersion in the measurements to dispersion in the measurand X, and dispersion in the measurements Y conditional on X. This is in line with the typical models employed in metrology and measurement theory.[1,17] Much of the literature of agreement studies, however, introduces κ statistics in the context of classifications and cross-tabulations, without reference to a measurand. By doing so, a mapping that is essentially a measurement is treated as merely a classification. The distinguishing characteristic of a measurement, is that it is a classification *aimed to reflect an empirical property of the subjects being measured* (cf. classical definitions of measurement).[18–20] Including this empirical property as an element of the statistical model, as is done in the model presented above, is not only more natural, it also allows one to separate the behaviour of the measurand (which is a characteristic of the subjects population) from the behaviour of the measurement errors (a characteristic of the classification procedure), and to state assumptions about both explicitly and separately. By defining κ indices in the context of classification and cross-tabulation, as is typically done in the literature, the assumptions about the measurand's behaviour are obscured. We speculate that the conception of agreement in the context of classification rather than measurement is one of the causes of many of the interpretation problems discussed in a later section.

Kraemer's[15] population model (her 'Case 1'), which is restricted to dichotomous classifications, allows a comparable distinction between 'characteristics of the population' and 'decision-making errors'. Contrary to our conceptualization, however, Kraemer sees the marginals $q(0)$ and $q(1)$ as population characteristics, rather than our $p(0)$ and $p(1)$, and this line of reasoning permeates, implicitly or explicitly, much of the literature on agreement. However, the $q(k)$ reflect both the distribution of true values in the subject population *and* the distribution of classification errors (per Equation (3)). We think it is better to regard the class prevalences $p(k)$ and the conditional probabilities $q(k|l)$ as the intrinsic characteristics of the subject population and measurement errors, respectively, and the marginal distribution given by the $q(k)$ as their combined consequence.

We discuss three problematic issues concerning agreement studies.

(1) Study design: nonrandom sampling
(2) Problems and errors related to nonuniform class prevalences
(3) Interpretation pitfalls

3 Study design: nonrandom sampling

The main error in the design of an agreement study based on κ, is that the sample is unrepresentative for the subject population. It is crucial to work with a representative sample. If one selects a sample in which the numbers of subjects in the different classes are not representative for the study population, $\hat{P}_A$ will be biased, because P_A depends on the class prevalences $p(l)$ (per Equation (4)), unless $P_A(l)$ is equal for all values of l. This bias will mostly be modest, depending on the differences between the $P_A(l)$ for different classes l. However, if one expresses the result in the form of κ, this bias is leveraged substantially by the rescaling based on $P_{A|C}$, and $\hat{\kappa}$ will be strongly biased even if the P_A (l) are equal for all l.

We illustrate the large bias that can result from unrepresentative sampling by a study that appeared in *The Lancet*, which researched the agreement in detecting the presence or absence of certain respiratory signs.[21] The authors do not state clearly how and from what population the patients were sampled. Most of the patients in the study had respiratory disorders and all patients had 'stable well-defined features and a definitive diagnosis confirmed by investigations', suggesting that the patients were not a random sample from the general population. If that is so, the estimated κ values are biased, or only representative for that specific population from which the sample was taken. We illustrate the possible magnitude of this bias from a numerical example. In the abovementioned study, a certain chest sign, an 'increased percussion note', has a $\hat{\kappa}$ value of 0.50. A total of 24 patients were each inspected by four physicians. The number of physicians that indicated an increased percussion note was 0 for 16 patients, 1 for five patients, 2 for one patient, 3 for none of the patients and 4 for two patients. This could correspond to the following statistical properties of the test procedure (values chosen for illustration, but not based on the original study): a prevalence of $p(1) = 0.10$, a specificity of $q(0|0) = 0.93$ and a sensitivity of $q(1|1) = 0.92$, which gives the reported value $\kappa = 0.50$. Now suppose that the sample is unrepresentative for the population that the researchers have in mind, because in fact the prevalence is not 0.10 but 0.01. Then, the population value of κ is only 0.10, and the study is likely to overestimate agreement by a factor 5. Clearly, κ depends heavily on prevalence and the sampling method, and it is crucial that the sample is representative. Therefore, when conducting an agreement study as in the paper on detecting respiratory signs, the population of subjects for

which the measurement procedure is intended, must be clearly defined, and when using κ, the sample of subjects must be a random sample from this subject population.

In the statistical model given above, the assumption was made that the m measurements Y_{i1}, $Y_{i2}, \ldots, Y_{im}$ are stochastically independent given the subject's true value X_i and, therefore, that X_i is the only factor affecting the stochastic properties of the measurements (the assumption of conditional independence). If the sample is representative, this assumption is not crucial. However, if the sample is unrepresentative, a violation of the assumption of conditional independence creates an even further bias.[22] In the study about detecting respiratory signs,[21] suppose that an increased percussion note is easier to detect for some patients than for others. This would be a violation of the conditional independence assumption. If the sample is unrepresentative in the sense that patients are overrepresented whose increased percussion note is relatively easy to detect, the expected value of $\hat{\kappa}$ will increase even beyond 0.50, leading to an even more serious overestimation.

The importance of a random sample is underappreciated in literature. The Food and Drug Administration[23] mentions that in studies evaluating diagnostic tests, the subjects should include the complete spectrum of patient characteristics, but does not mention that the sample of subjects should otherwise be representative for the study population. Several papers about agreement studies based on κ recommend a balanced sample, in which sample prevalences are uniform,[6,7,24,25] clearly in conflict with our observation that a representative sample is essential. Other misconceived sampling strategies include sampling subjects from the so-called *gray area*, i.e. subjects that are hard to judge, or sampling subjects such that roughly one-third is clearly positive, one-third clearly negative, and one-third hard to judge.

4 Problems and errors related to nonuniform class prevalences

Kraemer et al.[26] point out that in the case of scales with 3 or more classes, κ indices may obscure a poor consistency on two classes because it is averaged out with a possibly good consistency on the remaining classes. We think the situation becomes even more tricky when class prevalences are nonuniform (that is, $p(k) \gg p(l)$ for some $k \neq l$). Especially, for tests for disorders, this is typically the situation, because typically, $p(0) \gg p(1)$ (the fraction of subjects unaffected by the disorder is much larger than the fraction affected).

If class prevalences are nonuniform, κ and P_A by approximation only reflect the consistency in the most prevalent class. This is because the $P_A(l)$ are weighted by the prevalence of each of the classes as in Equation (4). This is not an error in itself, but it should be borne in mind in interpreting the κ index. For example, if a certain disease has a small prevalence ($p(1) \approx 0.0$), the P_A estimated from a random sample of diagnoses almost exclusively reflects the specificity $q(0|0)$:

$$\begin{aligned} P_A &= p(0)\big(q^2(0|0) + q^2(1|0)\big) + p(1)\big(q^2(0|1) + q^2(1|1)\big) \\ &\approx p(0)\big(q^2(0|0) + (1 - q(0|0))^2\big) \end{aligned}$$

and similarly for κ. Reporting the quality of the diagnostic procedure solely as a κ index, would fail to reflect the procedure's sensitivity $q(1|1)$, which is an equally important aspect of diagnostic quality. De Mast et al.[22] give a similar warning for pass/fail inspections in industry, where κ indices evaluate an inspection exclusively in terms of the producer's risk (the probability of a false rejection), ignoring the consumer's risk (a false acceptance).

Interpretation of κ becomes even more precarious if class prevalences are extremely nonuniform (one $p(l) > 0.95$). In such cases, the partial derivatives of κ with respect to the parameters $q(k|l)$

approach 0, while the partial derivative with respect to the $p(l)$ becomes very large (Figure 1). This implies that, in the case of extremely nonuniform class prevalences, κ responds strongly and rather one-sidedly to changes in class prevalences.

In summary, the κ index confounds various properties of the measurements and the class prevalences into a single number, and especially for nonuniform class prevalences is driven rather one-sidedly by either the $q(k|l)$ of the most prevalent class l, or (for extremely nonuniform prevalences) responds almost exclusively to the class prevalences themselves. As a consequence, it may be an oversimplification to reject or accept a measurement procedure on the basis of κ, following criteria[14,27,28] for values of κ. For example, Fleiss et al.[28] judge $\kappa < 0.40$ as poor, κ between 0.40 and 0.74 as fair to good and κ between 0.75 and 1.00 as excellent. To illustrate how uncritical application of such criteria leads to practically questionable decisions, consider a dichotomous test with specificity and sensitivity $q(0|0) = q(1|1) = 0.95$. For many practical applications, these error probabilities may be acceptable. However, if the prevalence $p(1) = 0.01$ (extremely nonuniform, but a common order of magnitude), then $\kappa = 0.14$, 'poor' according to the abovementioned criteria. Note that as a first-order (Taylor) approximation around these values,

$$\kappa \approx 0.14 + 12.3\,(p(1) - 0.01) + 2.6\,(q(0|0) - 0.95) + 0.3\,(q(1|1) - 0.95),$$

illustrating the fact that for such nonuniform prevalences, κ is largely driven by the prevalence (slope of 12.3), and that it is nearly insensitive to $q(1|1)$ (slope of 0.3).

Instead of the sensitivity and specificity, in different contexts, it might be more important to consider the positive and negative predictive values of a diagnosis.[17] Suppose a diagnosis has a positive predictive value $P(X = 1|Y = 1) = 0.8$ and a negative predictive value $P(X = 0|Y = 0) = 0.95$. Assuming a prevalence of $p(1) = 0.10$, the $\kappa = 0.39$ is poor according to

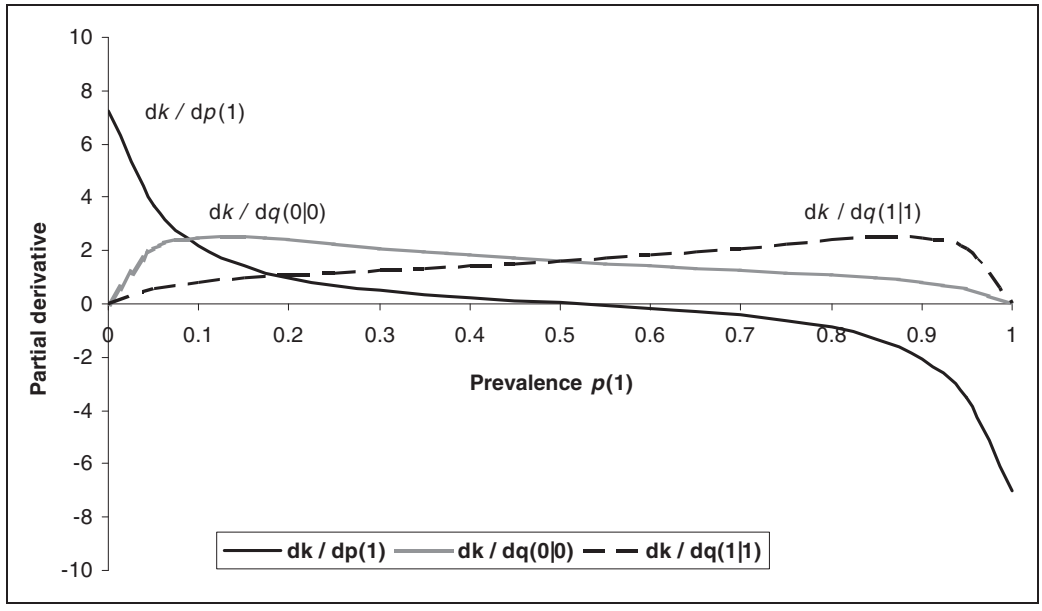

Figure 1. Partial derivatives of κ for various values of the prevalence $p(1)$, in the situation of a dichotomous scale, and assuming that $q(0|0) = q(1|1) = 0.95$.

Fleiss et al., but the predictive values may be acceptable for the application under study. A situation with worse positive predictive value (0.7 instead of 0.8) but higher prevalence (0.25) gives $\kappa = 0.48$, 'fair to good' according to the criteria.

Besides an interpretation problem, nonuniform class prevalences also result in a strong sensitivity of the estimator $\hat{\kappa}$ to sampling variations, and thus a large standard error of the estimator $\hat{\kappa}$. De Mast[11] gives an example for a two-point scale, involving 100 subjects rated by two appraisers, where changing a single data point reduces the estimated agreement $\hat{\kappa}$ from 1.0 to 0.66.

We conducted a simulation study to determine the standard error of $\hat{\kappa}$ statistics for dichotomous tests, where subjects are diagnosed as either negative (0) or positive (1). The standard error depends on the prevalence $p(1)$, specificity $q(0|0)$, sensitivity $q(1|1)$ and the sample size (the numbers of subjects n and appraisers m). For each combination of sample size and model parameters, 10 000 samples were created and for each sample, the $\hat{\kappa}$ statistic was computed, taking the sample standard deviation of these 10 000 realizations as the estimated standard error. With 99% confidence, the results in Tables 1 and 2 have a relative error of at most 2% (leading to an absolute error of 0.007 in extreme cases). Alternatively, the standard error could be approximated based on a multinomial distribution for the cell counts in a cross-tabulation.[26,29,30]

Tables 1 and 2 show that the standard error may be unacceptably large, especially if prevalence is below 0.10 and specificity above 0.90, a very common situation. In this situation, if the number of subjects is $n = 50$ and the number of appraisers is $m = 4$, the standard error is above 0.08 (and can get as large as 0.35). The same holds for $n = 100$ subjects and $m = 2$ appraisers, even for lower values of specificity. Bootstrapping shows that a 95% confidence interval on κ has a width larger than 0.35 in almost all these cases, which makes the estimate practically useless.

A potentially unacceptably large standard error of $\hat{\kappa}$ is a common problem in applications of κ in literature. A study[31] that appeared in *Endoscopy* is an example of this problem. It assesses inter-observer agreement of a certain type of endoscopy. Four endoscopists evaluated video sequences recorded during endoscopies of 51 patients with reflux symptoms. This is very close to the sample size discussed above. The prevalence is not reported in this article, but if it is low, the standard errors of the $\hat{\kappa}$ statistics may be extremely large. For example, the estimated $\hat{\kappa} = 0.36$ for the detection of 'Methylene blue positivity' might have a standard error of 0.18 if prevalence was 0.05, and the true agreement would be 'somewhere in between 0.09 and 0.79' (based on a bootstrapped 95% confidence interval, and taking for illustration that $q(0|0) = q(1|1) = 0.94$).

Another illustration of excessive standard errors is a paper[32] that appeared in the *Annual Review of Psychology*, which analysed 236 studies on youth psychotherapy. In order to assess the coding procedures used to categorize the studies, a 'master coder' and two students coded 30 randomly selected studies. Then, $\hat{\kappa}$ values were computed between the 'master coder' and each of the students and the mean of the two resulting $\hat{\kappa}$ statistics was taken. For this small sample size (30 subjects and 2 coders), if there are two categories, the standard error of $\hat{\kappa}$ is larger than 0.10 for almost all parameter values, which is unacceptably large. If prevalence (which is not reported in this article) is 0.05, the standard error can get as large as 0.31, again making the value of $\hat{\kappa}$ almost entirely uninformative. (The standard error of the mean of the two $\hat{\kappa}$ statistics is smaller, but still unacceptable.)

5 Interpretation pitfalls

The problematic interpretation of κ and its sometimes paradoxical behaviour have been much discussed in the literature.[6–12] The usual way κ is interpreted, is as the probability of agreement corrected for agreement by chance, in the sense that the value of κ is zero if P_A is equal to the probability of agreement of chance measurements. However, the concept of 'chance measurements'

Table 1. Standard errors of $\hat{\kappa}$ for $m = 2$

Standard error of kappa for $R = 10\,000$ (99% confidence interval maximum $\pm$ 0.007)

		spec.			
prev.	sens.	0.50	0.75	0.95	0.99
		$n = 30$, $m = 2$			
0.01	0.50	0.07	0.08	0.10	0.17
	0.75	0.08	0.08	0.15	0.27
	0.95	0.08	0.09	0.20	0.35
	0.99	0.08	0.09	0.21	0.36
0.10	0.50	0.07	0.08	0.15	0.21
	0.75	0.08	0.10	0.19	0.22
	0.95	0.08	0.12	0.21	0.21
	0.99	0.09	0.12	0.20	0.21
0.25	0.50	0.07	0.08	0.13	0.13
	0.75	0.08	0.10	0.12	0.12
	0.95	0.09	0.11	0.10	0.07
	0.99	0.09	0.11	0.10	0.06
0.50	0.50	0.08	0.08	0.10	0.10
	0.75	0.08	0.10	0.10	0.10
	0.95	0.10	0.10	0.08	0.06
	0.99	0.10	0.10	0.06	0.04
		$n = 50$, $m = 2$			
0.01	0.50	0.06	0.06	0.09	0.19
	0.75	0.06	0.06	0.13	0.28
	0.95	0.06	0.07	0.17	0.35
	0.99	0.06	0.07	0.18	0.37
0.10	0.50	0.06	0.06	0.12	0.16
	0.75	0.06	0.08	0.14	0.15
	0.95	0.07	0.09	0.15	0.12
	0.99	0.07	0.09	0.14	0.11
0.25	0.50	0.06	0.07	0.09	0.10
	0.75	0.06	0.08	0.09	0.09
	0.95	0.07	0.09	0.08	0.05
	0.99	0.07	0.09	0.07	0.04
0.50	0.50	0.06	0.06	0.08	0.08
	0.75	0.06	0.08	0.08	0.07
	0.95	0.08	0.08	0.06	0.05
	0.99	0.08	0.07	0.05	0.03
		$n = 100$, $m = 2$			
0.01	0.50	0.04	0.04	0.07	0.17
	0.75	0.04	0.04	0.10	0.25
	0.95	0.04	0.05	0.13	0.30
	0.99	0.04	0.05	0.14	0.32

(continued)

Table 1. Continued

		spec.			
prev.	sens.	0.50	0.75	0.95	0.99
0.10	0.50	0.04	0.04	0.09	0.11
	0.75	0.04	0.05	0.10	0.10
	0.95	0.05	0.06	0.10	0.07
	0.99	0.05	0.07	0.09	0.06
0.25	0.50	0.04	0.05	0.07	0.07
	0.75	0.04	0.06	0.07	0.06
	0.95	0.05	0.06	0.05	0.04
	0.99	0.05	0.06	0.05	0.03
0.50	0.50	0.04	0.05	0.05	0.05
	0.75	0.05	0.05	0.05	0.05
	0.95	0.05	0.05	0.04	0.03
	0.99	0.05	0.05	0.03	0.02
		$n = 200$, $m = 2$			
0.01	0.50	0.03	0.03	0.05	0.13
	0.75	0.03	0.03	0.07	0.18
	0.95	0.03	0.03	0.09	0.22
	0.99	0.03	0.03	0.10	0.23
0.10	0.50	0.03	0.03	0.06	0.07
	0.75	0.03	0.04	0.07	0.07
	0.95	0.03	0.05	0.07	0.05
	0.99	0.03	0.05	0.06	0.04
0.25	0.50	0.03	0.03	0.05	0.05
	0.75	0.03	0.04	0.05	0.04
	0.95	0.03	0.04	0.04	0.03
	0.99	0.04	0.04	0.03	0.02
0.50	0.50	0.03	0.03	0.04	0.04
	0.75	0.03	0.04	0.04	0.04
	0.95	0.04	0.04	0.03	0.02
	0.99	0.04	0.04	0.02	0.01

is too ambiguous to provide a well-defined zero point. Chance measurements are a nonexistent hypothetical concept, and therefore, anything said about their distribution is bound to be hopelessly arbitrary, and it does not make sense to argue about how appraisers would conduct 'chance measurements' in practice.

In fact, an analysis of the chance correction $P_{A|C}$ shows that its premises have implausible or even irreconcilable implications.[11] Chance measurements are assumed to have a distribution equal to the marginal distribution of the measurements under study (that is, $P(Z_{ij} = k) = q(k)$). It is hard to imagine why or by what sort of mechanism blind measurements would happen to have the same distribution as the measurement procedure under study. But besides being an implausible choice, a problematic consequence is that κ, thus interpreted, is unsuited to compare two different measurement procedures. For instance, Naranjo et al.[33] compare the agreement of a new procedure for classifying adverse drug reactions to current practice. However, the chance correction applied to the agreement of the new procedure is based on the marginal distribution of

Table 2. Standard errors of $\hat{\kappa}$ for $m = 4$

Standard error of kappa for $R = 10\,000$ (99% confidence interval maximum $\pm$ 0.007)					
		spec.			
prev.	sens.	0.50	0.75	0.95	0.99
		$n = 30$, $m = 4$			
0.01	0.50	0.07	0.08	0.10	0.17
	0.75	0.08	0.08	0.15	0.27
	0.95	0.08	0.09	0.20	0.35
	0.99	0.08	0.09	0.21	0.36
0.10	0.50	0.07	0.08	0.15	0.21
	0.75	0.08	0.10	0.19	0.22
	0.95	0.08	0.12	0.21	0.21
	0.99	0.09	0.12	0.20	0.21
0.25	0.50	0.07	0.08	0.13	0.13
	0.75	0.08	0.10	0.12	0.12
	0.95	0.09	0.11	0.10	0.07
	0.99	0.09	0.11	0.10	0.06
0.50	0.50	0.08	0.08	0.10	0.10
	0.75	0.08	0.10	0.10	0.10
	0.95	0.10	0.10	0.08	0.06
	0.99	0.10	0.10	0.06	0.04
		$n = 50$, $m = 4$			
0.01	0.50	0.06	0.06	0.09	0.19
	0.75	0.06	0.06	0.13	0.28
	0.95	0.06	0.07	0.17	0.35
	0.99	0.06	0.07	0.18	0.37
0.10	0.50	0.06	0.06	0.12	0.16
	0.75	0.06	0.08	0.14	0.15
	0.95	0.07	0.09	0.15	0.12
	0.99	0.07	0.09	0.14	0.11
0.25	0.50	0.06	0.07	0.09	0.10
	0.75	0.06	0.08	0.09	0.09
	0.95	0.07	0.09	0.08	0.05
	0.99	0.07	0.09	0.07	0.04
0.50	0.50	0.06	0.06	0.08	0.08
	0.75	0.06	0.08	0.08	0.07
	0.95	0.08	0.08	0.06	0.05
	0.99	0.08	0.07	0.05	0.03
		$n = 100$, $m = 4$			
0.01	0.50	0.04	0.04	0.07	0.17
	0.75	0.04	0.04	0.10	0.25
	0.95	0.04	0.05	0.13	0.30
	0.99	0.04	0.05	0.14	0.32

(continued)

Table 2. Continued

		spec.			
prev.	sens.	0.50	0.75	0.95	0.99
0.10	0.50	0.04	0.04	0.09	0.11
	0.75	0.04	0.05	0.10	0.10
	0.95	0.05	0.06	0.10	0.07
	0.99	0.05	0.07	0.09	0.06
0.25	0.50	0.04	0.05	0.07	0.07
	0.75	0.04	0.06	0.07	0.06
	0.95	0.05	0.06	0.05	0.04
	0.99	0.05	0.06	0.05	0.03
0.50	0.50	0.04	0.05	0.05	0.05
	0.75	0.05	0.05	0.05	0.05
	0.95	0.05	0.05	0.04	0.03
	0.99	0.05	0.05	0.03	0.02
		$n = 200$, $m = 4$			
0.01	0.50	0.03	0.03	0.05	0.13
	0.75	0.03	0.03	0.07	0.18
	0.95	0.03	0.03	0.09	0.22
	0.99	0.03	0.03	0.10	0.23
0.10	0.50	0.03	0.03	0.06	0.07
	0.75	0.03	0.04	0.07	0.07
	0.95	0.03	0.05	0.07	0.05
	0.99	0.03	0.05	0.06	0.04
0.25	0.50	0.03	0.03	0.05	0.05
	0.75	0.03	0.04	0.05	0.04
	0.95	0.03	0.04	0.04	0.03
	0.99	0.04	0.04	0.03	0.02
0.50	0.50	0.03	0.03	0.04	0.04
	0.75	0.03	0.04	0.04	0.04
	0.95	0.04	0.04	0.03	0.02
	0.99	0.04	0.04	0.02	0.01

the new procedure, whereas the chance correction applied to the agreement in current practice is based on the marginals in current practice. Thus, the two κ indices employ different chance corrections and therefore have values on scales with different zero points. The same comment holds for a study reported in *The Lancet*[34] that compares the agreement of two procedures for assessing the presence or absence of the ankle jerk in elderly people.

Another problematic consequence is that, on the one hand, chance measurements are conceived as blind (that is, uninformative about the measurand), but on the other hand, their distribution given by the $q(k)$ is related to the class prevalences $p(k)$ of the measurand (since the marginals $q(k)$ are related to the $p(k)$ via (3)); to the authors, these seem two irreconcilable implications.[11]

One possible solution is to use an unambiguous zero-point for correcting the raw P_A. De Mast and Van Wieringen[13] propose to define $P_{A|C}$ as the agreement of a maximally non-informative measurement system. Defining chance measurements as having a uniform distribution, that is, $P(Z_{ij}=k)=1/a$ for all k values, the resulting chance correction is $P_{A|C}^{\text{Unif}}=1/a$, and

$$\kappa^{\text{Unif}}=\frac{P_A-1/a}{1-1/a}.$$

This metric was proposed earlier by Bennett et al.[35] and advocated by Brennan and Prediger,[36] and others. It has at least two clear and unambiguous interpretations. First, $P_{A|C}=1/a$ is the lower bound for the probability of agreement attainable by measurement systems on a scale with a classes.[13] Second, chance measurements thus defined represent maximally non-informative measurements, in the information theoretic sense where information is defined as the negation of entropy, and the uniform distribution has maximal entropy.[37] Thus, κ^{Unif} is the probability of agreement in excess of minimal agreement on the given scale, or in excess of the agreement of maximally non-informative measurements.

An alternative solution is to interpret κ not as a measure of agreement corrected for agreement by chance, but as a measure of intraclass association. The problematic term $P_{A|C}$ is not interpreted in itself. Instead, the κ index can be shown to have the form of a measure of predictive association by rearranging its terms.[11] Let $\Delta_Z^G=1-\sum_{k=0}^{a-1}p_k^2$ be the Gini dispersion[38] of a categorical variable Z with a probability distribution $(p_0, p_1, \ldots, p_{a-1})$. Then

$$\kappa=1-\frac{1-\sum_{l=0}^{a-1}\left(p(l)\sum_{k=0}^{a-1}q^2(k|l)\right)}{1-\sum_{k=0}^{a-1}q^2(k)}=1-\frac{\Delta_{Y|X}^G}{\Delta_Y^G}.$$

The form $1-\Delta_{Y|X}/\Delta_Y$ on the right is the general expression of a coefficient of predictive association, where Δ can be any measure of dispersion,[39] and κ thus turns out to be a measure of association based on Gini's dispersion measure. Taking for Δ the entropy $\Delta_Z^E=-\sum_{k=0}^{a-1}p_k\log p_k$ instead of the Gini dispersion, one finds Theil's uncertainty coefficient,[11] which is thus a direct cousin of κ. It is also similar in form to the intraclass correlation coefficient (ICC) used to express the reliability of interval or ratio scale measurements:

$$ICC=Cor(Y_{i,1}, Y_{i,2})=1-\frac{\sigma_{Y_{ij}|X_i}^2}{\sigma_{Y_{ij}}^2}=1-\frac{\Delta_{Y|X}^V}{\Delta_Y^V},$$

with Δ^V now the variance instead of the Gini dispersion. Interpreted in this way, κ represents the association between two measurements $Y_{i,1}$ and $Y_{i,2}$ of the same subject i. Another analogue is the coefficient of determination R^2 in regression analysis, where $Y_{ij}=X_i+\varepsilon_{ij}$:

$$R^2=Cor^2(Y_{ij}, X_i)=\frac{\sigma_{X_i}^2}{\sigma_{Y_{ij}}^2}=1-\frac{\sigma_{Y_{ij}|X_i}^2}{\sigma_{Y_{ij}}^2}=1-\frac{\Delta_{Y|X}^V}{\Delta_Y^V}.$$

This gives the interpretation that κ represents the fraction of the total dispersion in the measurements Y_{ij} that can be attributed to dispersion in the measurands X_i, that is, as a measure

of reliability. Interpreting κ as a measure of predictive association, much of its paradoxical behaviour makes sense.

Kraemer et al.[26] dismiss the chance-corrected agreement interpretation as a historical curiosum, but focus on an interpretation as an *ICC*. Since their elaboration only holds for $a = 2$ classes, they recommend against the use of κ if $a \geq 3$. We think that our elaboration, based on the Gini dispersion, shows that κ *can* be interpreted as a reliability measure when $a \geq 3$.

Working with the interpretation of κ as a measure of association, it is important to be aware that such measures express measurement dispersion *in relation to a population of subjects* (and the class prevalences or distribution of the measurand in that population). If the same diagnostic test is applied in another population of subjects, with different class prevalences, then κ will be different. Consider, as an example, a dichotomous diagnostic test, with specificity and sensitivity $q(0|0) = q(1|1) = 0.95$. Depending on the prevalence $p(1)$ of the disorder, κ ranges from 0.00 to 0.81 in this case. In view of this fact that agreement is expressed in relation to prevalences in the subjects population, when expressing the results of an agreement study in terms of κ, it is crucial to define and delineate the study population of subjects to which it applies, and failure to do so makes the reported κ meaningless.

In Section 3, we have given an example of a paper about respiratory disorders[21] that does not clearly specify the relevant subject population. Another interesting example is Naranjo et al.,[33] who assess the reliability of classifying alleged adverse drug reactions (ADRs) by the probability that they were caused by drug therapy: definite, probable, possible or doubtful. The aim of the authors is to 'develop a simple method to assess the causality of ADRs in a variety of clinical situations'. They take a sample of 63 randomly selected cases published in a number of prestigious journals. The results therefore only apply to those ADRs that are published in prestigious journals, which cannot be assumed to be representative for the ADRs in clinical situations. The values of κ that they report are therefore meaningless for clinical situations.

6 Conclusion and recommendations

We conclude this article by listing our recommendations for agreement studies. Throughout this article, we have emphasized the role of the measurand, and our first recommendation is that a prerequisite for an agreement study is that the measurand is well (that is, clinically) defined. Without a clinical definition of the measurand, the whole concept of measurement error becomes meaningless. Second, it is important to clearly define and delineate the population of subjects in which the measurement procedure should be discriminating. Especially, if the κ index is used, the results are meaningless if the subject population is not well defined. Third, when using κ, the sample of subjects must be a random sample from the defined subject population; however, impractical that may be, since the estimation bias in $\hat{\kappa}$ may be substantial otherwise.

To evaluate the quality of the measurements, one may use the κ index if one wants a measure of reliability. We recommend against the interpretation as agreement corrected for agreement by chance, as the notion of chance agreement is too problematic and ambiguous. Instead, it can be interpreted as a reliability measure, much alike to the intraclass correlation coefficient used for interval and ratio scale measurements. However, if a gold standard is available to determine the measurand of each subject in the study, we recommend against the use of the κ index, and instead, propose to estimate the classification probabilities $q(k|l)$ conditional on the measurand; in the case of a dichotomous test, this amounts to establishing the sensitivity and specificity. Such studies are described in Pepe[17] and De Mast et al.,[11] and since they establish the intrinsic parameters of the measurement errors, we consider them more informative than agreement studies. If no gold standard is available, latent class or latent trait methods may be used to estimate the same parameters,[40]

although such methods critically depend on the viability of the conditional independence assumption.

Also, in applications with strongly nonuniform class prevalences, such as diagnoses of disorders with low prevalences, we recommend against using the κ index, because of three problematic properties in such cases. First, the standard errors of the estimates are typically unacceptably large. Second, the κ index is very sensitive to the class prevalences, and reflects prevalences more than it reflects the quality of the measurement procedure. Third, κ reflects rather one-sidedly the consistency on the most prevalent class. In such cases, numerical criteria for κ intended to evaluate a measurement procedure may be an oversimplification and lead to practically questionable decisions. One alternative is a sensitivity/specificity study mentioned above. Another option is to estimate κ^{Unif} as defined in an earlier section. It has a clear interpretation, and it does not suffer from the first two problems related to nonuniform class prevalences.

Conflict of interest statement

The authors declare that there is no conflict of interest.

References

1. ISO. *Guide to the expression of uncertainty in measurement*, 1st ed. Geneva, Switzerland: International Organization for Standardization, 1995.
2. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measure* 1960; **20**: 37–46.
3. Fleiss JL. Measuring nominal scale agreement among many raters. *Pychol Bull* 1971; **76**: 378–382.
4. Conger AJ. Integration and generalization of kappas for multiple raters. *Psychol Bull* 1980; **88**: 322–328.
5. Davies M and Fleiss JL. Measuring agreement for multinomial data. *Biometrics* 1982; **38**: 1047–1051.
6. Feinstein AR and Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990; **43**: 543–549.
7. Byrt T, Bishop J and Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol* 1993; **46**: 423–429.
8. Thompson WD and Walter SD. A reappraisal of the kappa coefficient. *J Clin Epidemiol* 1988; **41**: 949–958.
9. Uebersax JS. Diversity of decision-making models and the measurement of interrater agreement. *Psychol Bull* 1987; **101**: 140–146.
10. Grove WM, Andreasen NC, McDonald-Scott P, Keller MB and Shapiro RW. Reliability studies of psychiatric diagnosis: theory and practice. *Arch Gen Psych* 1981; **38**: 408–413.
11. De Mast J. Agreement and kappa type indices. *Am Stat* 2007; **61**: 148–153.
12. Warrens MJ. A formal proof of a paradox associated with Cohen's kappa. *J Classif* 2010; **27**: 322–332.
13. De Mast J and Van Wieringen W. Measurement system analysis for categorical data: agreement and kappa type indices. *J Qual Technol* 2007; **39**: 191–202.
14. Landis JR and Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–174.
15. Kraemer HC. Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika* 1979; **44**: 461–472.
16. Tanner MA and Young MA. Modeling agreement among raters. *J Am Stat Assoc* 1985; **80**: 175–180.
17. Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford, UK: Oxford University Press, 2003.
18. Lord FM and Novick MR. *Statistical theories of mental test scores*. Oxford, England: Addison-Wesley, 1968.
19. Allen MJ and Yen WM. *Introduction to measurement theory*. Monterey, CA: Brooks/Cole, 1979.
20. Wallsten TS. Measurement theory. In: Kotz S and Johnson N (eds) *Encyclopedia of statistical sciences*. Vol. 5, 8th ed. New York: Wiley, 1988.
21. Spiteri MA, Cook DG and Clarke SW. Reliability of eliciting physical signs in examination of the chest. *Lancet* 1988; **331**: 873–875.
22. De Mast J, Erdmann TP and Van Wieringen WN. Measurement system analysis for binary inspection: continuous versus dichotomous measurands. *J Qual Technol* 2011; **43**: 99–112.
23. Food and Drug Administration. *Guidance for industry and FDA staff – statistical guidance on reporting results from studies evaluating diagnostic tests*. Rockville, MD: U.S. Department of Health and Human Services, 2007.
24. Cicchetti DV and Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. *J of Clin Epidemiol* 1990; **43**: 551–558.
25. Hripcsak G and Heitjan DF. Measuring agreement in medical informatics reliability studies. *J Biomed Inform* 2002; **35**: 99–110.
26. Kraemer HC, Periyakoil VS and Noda A. Kappa coefficients in medical research. *Stat Med* 2002; **21**: 2109–2129.
27. Cicchetti DV and Sparrow SS. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *Am J Mental Defic* 1981; **86**: 127–137.
28. Fleiss JL, Levin B and Paik MC. *Statistical methods for rates and proportions*, 3rd ed. New York: John Wiley & Sons, 2003.
29. Fleiss JL, Cohen J and Everitt BS. Large sample standard errors of kappa and weighted kappa. *Psychol Bull* 1969; **72**: 323–337.

30. Bloch DA and Kraemer HC. 2 x 2 kappa coefficients: measures of agreement or association. *Biometrics* 1989; **45**: 269–287.
31. Meining A, Rösch T, Kiesslich R, Muders M, Sax F and Heldwein W. Inter- and intra-observer variability of magnification chromoendoscopy for detecting specialized intestinal metaplasia at the gastroesophageal junction. *Endoscopy* 2004; **36**: 160–164.
32. Weisz JR, Jensen Doss AJ and Hawley KM. Youth psychotherapy outcome research: a review and critique of the evidence base. *Ann Rev Psychol* 2005; **56**: 337–363.
33. Naranjo CA, Busto U, Sellers EM, Sandor P, Ruiz I, Roberts EA, et al. A method for estimating the probability of adverse drug reactions. *Clin Pharmacol Ther* 1981; **30**: 239–245.
34. O'Keefe ST, Smith T, Valacio R, Jack CIA, Playfer JR and Lye M. A comparison of two techniques for ankle jerk assessment in elderly subjects. *Lancet* 1994; **344**: 1619–1620.
35. Bennett EM, Alpert R and Goldstein AC. Communications through limited response questioning. *Public Opin Q* 1954; **18**: 303–308.
36. Brennan RL and Prediger DJ. Coefficient kappa: some uses, misuses, and alternatives. *Educ Psychol Measure* 1981; **41**: 687–699.
37. Berger T. Information theory and coding theory. In: Kotz S and Johnson N (eds) *Encyclopedia of statistical sciences*. Vol. 4, New York: Wiley, 1988, pp.124–141.
38. Gilula Z and Haberman SJ. Dispersion of categorical variables and penalty functions: derivation, estimation, and comparability. *J Am Stat Assoc* 1995; **90**: 1447–1452.
39. Hershberger SL and Fisher DG. Measures of association. In: Everitt B and Howell D (eds) *Encyclopedia of statistics in behavioral science*. Vol. 3, Chichester: Wiley, 2005, pp.1183–1192.
40. Van Wieringen WN and De Mast J. Measurement system analysis for binary data. *Technometrics* 2008; **50**: 468–478.