View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Biomedical Informatics 35 (2002) 99–110

Journal of Biomedical Informatics

www.academicpress.com

Measuring agreement in medical informatics reliability studies

George Hripcsak[a,*] and Daniel F. Heitjan[b]

[a] *Department of Medical Informatics, Columbia University, 622 West 168th Street, VC5, New York, NY 10032, USA*
[b] *Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, USA*

Received 8 May 2002

Abstract

Agreement measures are used frequently in reliability studies that involve categorical data. Simple measures like observed agreement and specific agreement can reveal a good deal about the sample. Chance-corrected agreement in the form of the kappa statistic is used frequently based on its correspondence to an intraclass correlation coefficient and the ease of calculating it, but its magnitude depends on the tasks and categories in the experiment. It is helpful to separate the components of disagreement when the goal is to improve the reliability of an instrument or of the raters. Approaches based on modeling the decision making process can be helpful here, including tetrachoric correlation, polychoric correlation, latent trait models, and latent class models. Decision making models can also be used to better understand the behavior of different agreement metrics. For example, if the observed prevalence of responses in one of two available categories is low, then there is insufficient information in the sample to judge raters' ability to discriminate cases, and kappa may underestimate the true agreement and observed agreement may overestimate it.

© 2002 Elsevier Science (USA). All rights reserved.

Keywords: Agreement; Reliability; Kappa; Latent structure analysis; Tetrachoric correlation; Prevalence

1. Introduction

Medical informatics researchers often employ agreement measures to quantify the similarity of two or more raters in responding to a series of tasks (e.g., assessing cases) [1–7]. For example, a researcher may wish to assess the performance of a decision support system in an area where there is no simple way to know the true state of the patient or the optimal therapeutic plan. The researcher may therefore enlist the help of domain experts (referred to here as the *raters*) to create a reference standard that can be compared to the output of the system [8]. Depending on the skill of the raters and the difficulty of the cases, however, the generated reference standard may or may not be appropriate. The researcher may therefore carry out a *measurement study* [1] to assess the quality of the reference standard before actually using the reference standard in a *demonstration study* [1] (in this example, calculating the performance of the system is the demonstration study).

One measure of the trustworthiness of a reference standard is the reliability (reproducibility, precision) of the raters' responses [1,9]. For categorical responses, agreement can be used as a measure of reliability [9,10]. If there is little agreement among the raters, then their responses are unreliable and the quality of the reference standard is suspect. (This use of agreement depends upon the properties of the data, such as the balance of the sample, and it is discussed in greater detail below.)

The goal of this paper is to review the common alternatives for measuring agreement, to discuss and demonstrate the relative strengths and weaknesses of the measures, and to provide high-level recommendations for when the measures should be used. The paper focuses on examples with two raters, but the described methods are extensible to multiple raters (and references are given).

2. Agreement metrics

2.1. Observed agreement

A large number of agreement measures have been suggested in the literature. A number of the simpler ones

[*] Corresponding author. Fax: 1-212-305-3302.
E-mail address: hripcsak@columbia.edu (G. Hripcsak).

1532-0464/02/$ - see front matter © 2002 Elsevier Science (USA). All rights reserved.
PII: S1532-0464(02)00500-2

are covered in reviews by Fleiss [10,11]. Three stand out in popularity: observed agreement [11,12], specific agreement [10,11], and kappa, a form of chance-corrected agreement [11,13].

Observed agreement (simple agreement, raw agreement) is the portion of cases for which the raters agree. If there are two raters and responses are dichotomous (say, positive and negative), then given the two-by-two contingency table in Table 1, observed agreement (A_o) is defined as follows:

$$A_o = \frac{a+d}{a+b+c+d}.$$

The observed agreement for the example in Table 2 is $.73 = (15+26)/56$. The extension to more than two categories is straightforward. Observed agreement is simply the proportion of cases for which the two raters agree. If the data are presented in an M-by-M contingency table, where M is the number of categories, then observed agreement is the sum along the diagonal divided by the total number of cases.

The extension to more than two raters is usually taken as mean pair-wise agreement [11], which is the average agreement across all possible pairs of raters. An alternative compares each rater to the majority opinion of the others [11].

Observed agreement can be misleading, however, because a certain amount of agreement is expected by chance. If the prior probability of a rater responding positive or negative is 0.5, then if both raters guess at random, their expected agreement is 0.5. Thus, what might be interpreted as a significant agreement is achieved with no real effort.

Observed agreement also lumps together the agreement on each of the categories when in fact the agreement may differ for each category. For example, if two raters are asked to judge the presence of a rare disease, the fact that they agree on the more frequently occurring negative cases is of little comfort if there is strong disagreement about which few cases are positive. Table 3 shows such an example. The observed agreement of 0.88 does not reflect the disagreement on the positive cases.

Table 1
Two-by-two contingency table

Rater A's judgment	Rater B's judgment		Total
	Positive	Negative	
Positive	a	b	$a+b$
Negative	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

Table 2
Example of a two-by-two contingency table

Rater A's judgment	Rater B's judgment		Total
	Positive	Negative	
Positive	15	6	21
Negative	9	26	35
Total	24	32	56

Table 3
Sample two-by-two contingency table for mediocre ability to diagnose a rare disease

Rater A's judgment	Rater B's judgment		Total
	Positive	Negative	
Positive	4	6	10
Negative	8	102	110
Total	12	108	120

2.2. Specific agreement

Specific agreement quantifies the degree of agreement for each of the categories separately [10,14]. In the dichotomous case, one can calculate positive and negative specific agreement. They are defined as the proportion of cases in a category for which the raters agreed. For the contingency table in Table 1, positive specific agreement (p_{pos}) and negative specific agreement (p_{neg}) are as follows:

$$p_{\text{pos}} = \frac{2a}{2a+b+c},$$

$$p_{\text{neg}} = \frac{2d}{b+c+2d}.$$

For the example in Table 2, positive specific agreement is .67 and negative specific agreement is .78. For the example in Table 3, positive specific agreement is .36 and negative specific agreement is .94, which reflects the discord in agreement on positive and negative cases.

2.3. Kappa

Another metric that has been used in reliability studies is kappa [11,13]. Kappa (κ) is defined in terms of observed agreement (A_o) and agreement expected by chance (A_e):

$$\kappa = \frac{A_o - A_e}{1 - A_e}.$$

This operation normalizes the scale so that chance agreement becomes zero. The agreement expected by chance is calculated in the same way as the observed agreement, except that the observed values in the cells are replaced with their expected values [1], which are based on the observed proportion of responses in each category. The expected value in cell a, $E[a]$, is $(a+b)(a+c)/(a+b+c+d)$ and the expected value in cell d, $E[d]$, is $(b+d)(c+d)/(a+b+c+d)$. The sum of the observed values $(a+b+c+d)$ equals the sum of the expected values. The agreement expected by chance is then as follows:

$$A_e = \frac{E[a]+E[d]}{a+b+c+d}.$$

The agreement expected by chance for the example in Table 2 is $.52 = (9+20)/56$. Kappa for Table 2 is then $.44 = (.73-.52)/(1-.52)$. For the example in Table 3, kappa is 0.3, far lower than the observed agreement (0.88). The frequent agreement on negative cases, which results in a high observed agreement, is ascribed to chance when kappa is calculated (chance agreement is based on the observed prevalences of each category), leaving the poor agreement on positive cases. This effect is described in greater detail below.

For a two-by-two contingency table (Table 1), the formulae for kappa can be combined and simplified to produce the following form that can be calculated directly from the table [10]:

$$\kappa = \frac{2(ad-bc)}{(a+c)(c+d)+(b+d)(a+b)}.$$

The procedure is similar for more than two categories. Observed agreement for multiple categories is calculated as described above. Chance agreement is calculated using the same approach but using the expected values instead of the observed values in the cells. Kappa can also be extended for multiple raters and multiple categories based on pair-wise agreement [9,15].

A weighted version of kappa has also been defined [10,16] which allows one to assign different penalties to different mismatches. For example, if the data are ordinal then one can assign a greater penalty when raters' responses are further apart on the ordinal scale. It is also possible to define kappa when different groups of raters assess each case [17], even when the number of raters varies per case [10].

2.4. Relation of kappa to intraclass correlation coefficients

Kappa is related to intraclass correlation coefficients and behaves similarly to them [10,11]. An intraclass correlation coefficient [9,18] can be defined for continuous data as the proportion of the overall variance in a sample that represents the differences of interest (real differences between cases, denoted "true variance") as opposed to noise such as inter-rater error ("error variance"):

$$\text{intraclass correlation coefficient} = \frac{\text{true variance}}{\text{true variance} + \text{error variance}}.$$

A higher proportion implies more reliable raters. In reliability studies, an intraclass correlation coefficient of 0.7 is commonly used as a threshold of "sufficient reliability."

From this definition, a formula to calculate the intraclass correlation coefficient can be defined in terms of mean square estimates from an analysis of variance [18]. A common form is

$$\text{interaclass correlation coefficient} = \frac{ms_{\text{between}} - ms_{\text{within}}}{ms_{\text{between}} + (k-1)ms_{\text{within}}},$$

where ms_{between} is the mean-square estimate of the between-subjects variance and ms_{within} is the mean-square estimate of the within-subjects variance for an analysis of variance on a case by rater experiment, and k is the number of raters. In this example, the "subjects" are the cases. There are actually several forms of intraclass correlation coefficients depending on whether the same raters judge each case, whether these raters are the only raters of interest or whether they represent a sample of the judges of interest, and whether one rater or all raters will be judging each case in the demonstration study [18].

Kappa applies to categorical data, but it can be defined in terms of sample mean squares and a correspondence can be drawn between it and an intraclass correlation coefficient [10,11]. Thus kappa can be seen as a natural measure of reliability. Fleiss [11] exploits the correspondence between kappa and intraclass correlation coefficients to compare various alternative measures of agreement and to argue that kappa is the appropriate one in many circumstances.

2.5. Interpreting levels of kappa

Perfect agreement is indicated by a kappa of one, and pure chance is indicated by a kappa of zero. Negative kappa indicates disagreement greater than that expected by chance. Unfortunately, intermediate levels of kappa between zero and one cannot be interpreted consistently. Fleiss [10] and Koch and Landis [10,19] have suggested 0.4 as the minimum that suggests fair to good agreement with values greater than 0.75 suggesting excellent agreement. In the non-medical classification literature, values greater than 0.67, 0.7, or 0.8 have been recommended [20]. In fact, the interpretation of these levels relies heavily on the tasks and categories, the purpose of the measurement, and the definition of chance, so such guidelines are deceptive and should probably not be used.

For example, if two experiments have identical tasks with ordinal scales except that one experiment has more levels on its scale, then kappa will be lower in the experiment with more levels even though the raters are identical [21]. In another example, as the prevalence of cases in each category becomes very high or very low, kappa approaches zero even if the raters are reliable [12,22,23] (this is covered further below).

The level of kappa that represents sufficient reliability to conduct an experiment depends on the goal. An experiment in which a system's answers for each case will be analyzed in detail (potentially resulting in system modifications) requires raters that are several times more reliable (a reliability coefficient of 0.95 has been suggested) than those in an experiment in which only overall system performance will be estimated [24,25].

Dunn [9] shows examples in which different choices of kappa and different definitions of kappa can result in values ranging from 0.21 to 0.80 for the same underlying data set. Most of the variation could be attributed to treating ordinal data in different ways: looking for an exact mach on the ordinal categories; binning categories to produce dichotomous results; or weighting mismatches by how far apart the raters' responses were for a given case (in terms of the number of categories). The right choice depends upon how the information will be used: is an exact match actually required; will the results actually be binned in the demonstration study; or are large mismatches (e.g., for a given question, one rater responds "strongly agree" and the other rater responds "strongly disagree") more severe than small mismatches (e.g., one rater responds "strongly agree" and another rater responds "agree"). The latter is likely to be the case in most experiments. Nevertheless, the point is that this choice can affect the level of kappa greatly.

3. Modeling decision making

Furthermore, kappa's correction for chance may be deceptive. The correction really depends upon one's model of how decisions are made. For example, there may be some underlying and hidden continuous trait (a *latent trait*) that the raters are really judging, such as the degree of evidence supporting a diagnosis. Raters assess the degree of evidence in each case with some amount of variability or error. Given their estimated degree of evidence, raters' respond with one of the categories (dichotomous or ordinal) based on some internal threshold.

Two raters may agree perfectly on the underlying trait but, due to a difference in threshold, their category assignments may match on few cases (one may be higher than the other most of the time). Kappa (more appropriately, weighted kappa) will be low here, implying that raters' responses are similar to chance assignment when in fact, they have perfect agreement on the underlying trait. Inspection of an M-by-M contingency table (for M ordinal categories and two raters) would quickly reveal that responses are not at all chance-like, with most of the responses clustered along a line parallel to but off the main diagonal.

This latter example points out the importance of modeling the decision making process. For example, a difference in thresholds is more easily rectified through training and feedback to the raters than many other sources of disagreement. Under the assumption that there is a latent trait, one can calculate the tetrachoric correlation [26] (for dichotomous data with two raters), polychoric correlation (for ordinal data with two raters), or latent trait models [9,27] (for dichotomous or ordinal data with more than two raters). There is no simple formula for these metrics; they must be estimated iteratively by computer. The output from these models includes an estimate of the correlation among the raters on the latent trait and an estimate of their relative thresholds.

Table 4 shows an extreme example to illustrate the behavior of these models. In all the disagreements, rater A calls a case negative and rater B calls it positive. Observed agreement is 0.5 and kappa is 0.2, both signifying poor agreement in this sample. Tetrachoric correlation is 1 (or close to it, depending on how it is estimated) with thresholds of −0.67 and 0.67 for the two raters. All the disagreement can be explained by a difference in threshold rather than disagreement on the underlying trait.

Latent trait and similar models usually assume a normal distribution for the trait and errors. Unfortunately, there is no way to test the assumptions of the model with a two-by-two table, but assumptions can be tested when there are more than two categories or more than two raters. Nevertheless, as with kappa, estimates of the magnitude of agreement must be interpreted with caution.

A different model assumes that data fall into two or more underlying and hidden classes (*latent classes*) and that raters have varying degrees of ability to map from the latent classes to the correct responses [9,27–29]. These latent class models are appropriate for dichotomous data, nominal data, and, in certain cases, ordinal data when there are at least three raters. In the simplest example, there are two classes: truly positive cases and truly negative cases. Raters' ability to discriminate cases can be defined by the probability of calling positive cases positive (sensitivity) and probability of calling negative cases negative (specificity). This model assumes that decisions are independent conditional on the true state (latent class). This model is usually overly simplistic, as there are usually some cases that are easier and some

Table 4
Sample two-by-two contingency table for a difference in threshold

Rater A's judgment	Rater B's judgment		Total
	Positive	Negative	
Positive	25	0	25
Negative	50	25	75
Total	75	25	100

cases that are harder, for which raters are more or less likely to agree. By adding classes (e.g., adding intermediate to positive and negative), one can model the varying difficulty of cases. In fact, by adding a number of classes, one can mimic a latent trait model. Such models can be used not just to quantify agreement, but also to understand the sources of disagreement in detail. For example, Espeland and Handelman [29] analyzed decision making and sources of disagreement in radiographic diagnosis of dental caries.

4. Effect of rater discrimination on agreement in a balanced sample

Researchers often quantify agreement to infer the quality and consistency of the raters. Assuming raters are independent, they will agree based on properly discriminating among the categories. Raters with better discrimination should have better agreement. In this section, we investigate the relation of agreement to rater discrimination to give a better feel for how different levels of agreement relate to different levels of discrimination [23].

For this discussion, we will use the simple model described above, in which cases belong to one of two classes, positive and negative. Raters' responses are conditionally independent given the class of a case. We chose this simple model because it illustrates the points clearly. In practice, most problems do include cases that are more or less difficult, warranting a model with more than two latent classes or a latent trait.

Table 5 shows the agreement for a series of scenarios, which are described briefly in the *description* column. In each scenario, there are two raters and responses are dichotomous. The data are generated from the following simulations. *Rater sensitivity* (second column) equals the raters' probability of assigning a positive response when the correct response for a case is positive, and *rater specificity* equals the raters' probability of assigning a negative response when the correct response for a case is negative. Taken together, these numbers describe the raters' ability to discriminate positive from negative cases, which is quantified in the next column as the mean area under the receiver operating characteristic curve (*mean ROC area*), calculated by the non-parametric method for when there is a single point [30]. The *underlying sample prevalence* is the prevalence of truly positive cases. A "balanced sample," as noted in the description column, is defined here as one in which the underlying sample prevalence is near .5. The information in columns 2–5 is not normally available to the researcher. It is used here to generate illustrative scenarios.

The next group of columns $(P\{a\}, P\{b\}, P\{c\}, P\{d\})$ show the expected proportion of counts in cells *a*, *b*, *c*, and *d* in Table 1, given the information in columns 2–5. Thus, in scenario 2, a contingency table might contain 4804, 196, 196, and 4804 for cells *a*–*d*, respectively (this is in fact the exact expected ratio among the cells for this scenario). Of course, in an actual experiment, the measured counts per cell would vary by chance. The underlying sample prevalence differs from the expected observed prevalence $(P\{a\} + P\{b\}/2 + P\{c\}/2)$ because imbalance in raters' ability to judge positive and negative cases (as in scenario 4) will shift the relative numbers of cases in the cells of the contingency table.

The agreement metrics are then reported in the following columns. They are calculated from the exact expected proportions per cell for each scenario. Observed agreement, specific agreement, kappa, and tetrachoric correlation (the correlation coefficient and the two rater thresholds). While tetrachoric correlation is calculated for these scenarios, it is intended for a latent trait model, not a model with two latent classes. If the decision making process truly follows a two class model, then one would not normally use tetrachoric correlation.

In scenario 1, both raters have perfect ability to discriminate between positive and negative cases. This is reflected by perfect observed agreement, specific agreement, kappa, and tetrachoric correlation. In scenario 2, both raters have excellent ability to discriminate between positive and negative cases. The agreement metrics remain high. Tetrachoric correlation is near one, and the thresholds are zero because the sample is balanced. In scenario 3, the raters have lower but still reasonable discrimination, and the agreement metrics are correspondingly lower.

The next four scenarios illustrate the effect of different patterns of sensitivity and specificity on agreement, given similar overall ability to discriminate (ROC area equal .87). In scenario 4, the raters have a mediocre ability to discriminate positive from negative cases and the sensitivity and specificity are equal, as in the earlier scenarios. As expected, the agreement metrics are lower than that for the earlier scenarios.

In scenario 5, raters have the same overall ability to discriminate cases as in scenario 4, but their sensitivity (.50) is far worse than their specificity (.99). A spread between sensitivity and specificity is a common finding in medical informatics. For example, in diagnosing a difficult disease, experts can easily eliminate the vast majority of negative cases because most of them may be completely unrelated to the disease under study (e.g., few overlapping symptoms) depending on the population. Distinguishing one complex disease from one or two other very similar diseases may be difficult, however, leading to lowered sensitivity in assigning the positive cases. A wide spread between sensitivity and specificity was found in experts' reading of radiology reports [7], for example.

Despite a balanced underlying sample, the difference between sensitivity and specificity in scenario 5 leads to

Table 5
Agreement for a series of balanced scenarios

Description	Rater sensitivity	Rater specificity	Mean ROC area	Underlying sample prevalence	$P\{a\}$[a]	$P\{b\}$[a]	$P\{c\}$[a]	$P\{d\}$[a]	Observed agreement	Positive specific agreement	Negative specific agreement	Kappa	Tetrachoric correlation (correlation; thresholds)
1. Perfect discrimination, balanced sample	1.00	1.00	1.00	.50	.500	.000	.000	.500	1.00	1.00	1.00	1.00	1.00; .00, .00
2. Excellent discrimination, balanced sample	.98	.98	.99	.50	.480	.020	.020	.480	.96	.96	.96	.92	1.00; .00, .00
3. Good discrimination, balanced sample	.92	.92	.96	.50	.426	.074	.074	.426	.85	.85	.85	.71	.89; .00, .00
4. Mediocre discrimination, balanced sample	.79	.79	.87	.50	.334	.166	.166	.334	.67	.67	.67	.34	.50; .00, .00
5. Mediocre discrimination (sens. $\ll$ spec.), balanced sample	.50	.99	.87	.50	.125	.130	.130	.615	.74	.49	.83	.32	.51; −.66, −.66
6. Mediocre discrimination (rater thresholds differ), balanced sample	.90, .67[b]	.67, .90[b]	.87	.50	.318	.067	.297	.318	.64	.64	.64	.31	.55; .29, −.29
7. Mediocre discrimination (one rater better overall), balanced sample	.90, .70[b]	.90, .70[b]	.87	.50	.330	.170	.170	.330	.66	.66	.66	.32	.48; .00, .00
8. No discrimination, balanced sample	.50	.50	.50	.50	.250	.250	.250	.250	.50	.50	.50	.00	.00; .00, .00

[a] $P\{a\} + P\{b\} + P\{c\} + P\{d\}$ may not sum to 1.000 due to rounding.

[b] For the first rater and second rater, respectively.

an imbalance between $P\{a\}$ and $P\{d\}$. Thus the apparent prevalence of positive cases $(P\{a\} + P\{b\}/2 + P\{c\}/2)$ is lower than .50. Observed agreement is higher for this scenario than for scenario 4, whereas kappa is similar to that for scenario 4. Specific agreement on positive cases is lower than that for negative cases, reflecting the shift in apparent in prevalence; this is covered in the following section. Tetrachoric correlation is similar to that of scenario 4, but the thresholds are now both negative, reflecting the apparent imbalance of cases (cell d greater than cell a).

In scenario 6, raters have different thresholds of deciding whether a case is positive or negative, which manifests itself as one rater being more sensitive and the other rater being more specific, but both having equal discriminating power. This leads to an imbalance between $P\{b\}$ and $P\{c\}$. Observed agreement, specific agreement, and kappa are similar to those in scenario 4. Tetrachoric correlation is higher than that for scenario 4, and the thresholds are of opposite sign. Thus some of the disagreement is attributed to a difference in threshold rather than poor correlation on some postulated latent trait.

In scenario 7, one rater is superior to the other, having both better sensitivity and specificity, although the combined ability to discriminate cases (in terms of mean ROC area) is similar to that of scenarios 4–6. All the agreement metrics are similar to those of scenario 4.

In scenario 8, the raters have no ability to discriminate cases, and all the agreement metrics reflect chance agreement.

In summary, all the agreement metrics decreased from perfect agreement (1) to chance agreement (0 or 0.5 in these examples) as raters' ability to discriminate cases decreased. Different patterns of sensitivity and specificity led to similar levels of measured agreement as long as overall discrimination was kept constant (although kappa was more stable than observed agreement in scenario 5). Kappa has the property that it, like a reliability coefficient, is zero when there is no discrimination (no correlation or reliability). This does not imply, however, that 0.7 is a magical number below which a demonstration study should not be conducted or above which a demonstration study can be conducted. Instead the scale is continuous: higher kappa implies more reliable judges, less unwanted variance, and a smaller sample size needed to detect a given effect.

Choosing between kappa and tetrachoric correlation (which is also zero for no discrimination) depends on the researcher's model of decision making. Tetrachoric correlation would not be appropriate for the two class model in these scenarios, but if a latent trait is suspected and if there is a need to distinguish correlation from threshold differences (e.g., to train raters and improve reliability), then tetrachoric correlation is appropriate.

5. Effect of prevalence (unbalanced sample) on agreement

Despite kappa's popularity, several authors have commented on its apparent shortcomings [12,14,22, 23,31]. In one formulation of the problem [12], kappa is sensitive to prevalence. If the prevalence of positive responses is near one or zero, then kappa may be close to zero despite high observed agreement.

The problem is illustrated in scenarios 9 and 10 in Table 6. Scenario 9 (copied from scenario 2 in Table 5) has a balanced sample. It demonstrates high agreement when the raters have excellent power to discriminate cases.

In scenario 10, the raters have the same excellent ability to discriminate cases, but the underlying prevalence of positive cases in the sample is very low (.01). This situation occurs frequently in medical informatics. For example, if raters are asked to judge the correctness of an information system's output and if that system itself is accurate, then the prevalence of incorrect cases will be low (an incorrect case is here defined as a "positive"). In another example, raters may be asked to judge the presence of a rare disease; there may be too few cases available to create a balanced sample of reasonable size.

In scenario 10, observed agreement (.96) is as high as that in scenario 9, but kappa (.32) falls to a level consistent with mediocre rater discrimination. It seems that studying an accurate system or a rare disease has made the raters look worse than they really are.

Whitehurst [31] argues that the dependence on prevalence is an unfortunate property of intraclass correlation coefficients (the argument applies also to kappa, which behaves similarly). He uses the example of measuring agreement among manuscript reviewers in the social sciences. The prevalence of high quality manuscripts is low, so the measured intraclass correlation coefficient is low (.38 in one example) despite high observed agreement (.80). He argues that the low prevalence unfairly penalizes reviewers because they were presented with an unbalanced sample.

Whitehurst recommends the use of Finn's r (last column of Table 6), an agreement metric that defines chance agreement as equal distribution in all categories. That is, if left to chance alone, raters would have put an equal number of responses in each category (e.g., positive and negative).

In scenario 9, kappa and Finn's r (Table 5) are equal because the observed distribution has an equal number of positive and negative responses. In scenario 10, however, the observed distribution has many more negative responses. Whereas Finn's r attributes the strong agreement on negative responses ($P\{d\}$ is near one) to good rater agreement, kappa attributes the high $P\{d\}$ to chance agreement. The effect is striking: kappa (.32) signifies mediocre agreement while Finn's r (.92)

Table 6
Agreement for a series of unbalanced scenarios

Description	Rater sensitivity	Rater specificity	Mean ROC area	Underlying sample prevalence	$P\{a\}$[a]	$P\{b\}$[a]	$P\{c\}$[a]	$P\{d\}$[a]	Observed agreement	Positive specific agreement	Negative specific agreement	Kappa	Tetrachoric correlation (correlation; thresholds)	Finn's r
9. (from scenario 2 Table 5) Excellent discrimination, balanced sample	.98	.98	.99	.50	.480	.020	.020	.480	.96	.96	.96	.92	1.00; .00, .00	.92
10. Excellent discrimination, unbalanced sample	.98	.98	.99	.01	.010	.020	.020	.951	.96	.34	.98	.32	.69; −1.88, −1.88	.92
11. Mediocre discrimination (sens. $\ll$ spec.), unbalanced sample	.50	.99	.87	.04	.010	.020	.020	.951	.96	.34	.98	.32	.69; −1.88, −1.88	.92
12. (from scenario 5 Table 5) Mediocre discrimination (sens. $\ll$ spec.), balanced sample (from Table 5)	.50	.99	.87	.50	.125	.130	.130	.615	.74	.49	.83	.32	.51; −.66, −.66	.48

[a] $P\{a\} + P\{b\} + P\{c\} + P\{d\}$ may not sum to 1.000 due to rounding.

signifies strong agreement. It appears that Finn's r better captures these raters' excellent discrimination.

Scenario 11 demonstrates the danger of using Finn's r in such circumstances. The raters have only mediocre discrimination, identical to that in scenario 12 (copied from scenario 5 in Table 5). The scenario 11 sample is highly unbalanced, resulting in cell frequencies ($P\{a\}-P\{d\}$) that are almost identical to those in scenario 10. Thus in scenario 11, kappa (.32) better represents the true discrimination of the raters than does Finn's r (.92) or the observed agreement (.96).

The problem is not that kappa is too low when the sample is unbalanced. The problem is that a severely unbalanced sample does not contain sufficient information to distinguish excellent raters from mediocre ones. Use of observed agreement or Finn's r might reassure a researcher that all is well when in fact the raters are poor. Kappa, while it might be inappropriately low, at least alerts the researcher that there is some problem in the raters, in the sample, or in both. Note that this is *not* an issue of sample size but instead of relative proportion. Even if there were 1,000,000 cases in scenario 9 with 10,000 in cell a, kappa would be .34.

Cicchetti and Feinstein [14] recommend the use of specific agreement in an unbalanced sample. In scenarios 10 and 11, the low positive specific agreement (.34 from Table 5) reflects the fact that agreement on the positive responses was low. The high negative specific agreement (.98 from Table 5) reflects the fact that agreement on the negative responses was high. One can conclude that the raters must have had a reasonable ability to discriminate negative cases (thus the high specificity in both scenarios). One cannot conclude what the raters' discrimination is for positive cases, however. Thus these two metrics highlight that there is a problem, and they point out what can be known (good agreement on negative cases), but they do not help distinguish scenario 10 from scenario 11. The sample simply lacks the information.

When there are more than two categories, specific agreement can highlight difficulties in particular categories, whereas kappa can only give an overall measure. For example, if there are three categories and one category is rare but agreement is good on the common categories, then kappa will be high but specific agreement will be low for the rare category, alerting the researcher to the issue.

Byrt and colleagues [12] go through the exercise of defining a prevalence-corrected version of kappa, and then point out that it is simply a rescaled version of the observed agreement (it goes from −1 to 1 instead of 0 to 1). They conclude that no index can capture the desired information (the discrimination of the raters) in this setting and that one should therefore report not just agreement but also a quantitative indicator of prevalence (that is, the balance of the sample). The researcher must then decide whether or not to draw a conclusion from the data.

All of the authors that commented on kappa agreed that ideally, one would use a balanced sample and avoid the issue of very low or very high prevalence [12,14,22,31]. Sometimes an inappropriate sample arises because of a failure on the part of the researcher to separate the measurement study from the demonstration study. For example, in the study of system correctness, one would want an even distribution of correct and incorrect cases to carry out a measurement study to assess the reliability of raters in judging the correctness of the system (as in scenario 9). If it is a good system, the demonstration study will have a low prevalence of incorrect cases (as in scenario 10). The error occurs when the data from the demonstration study are used in the measurement study. The measurement study should use a more balanced sample. That is, the sample for estimating rater reliability (agreement) should be based on what the researcher is trying to distinguish, not whatever sample happens to be created in the demonstration study.

The same authors recognize, however, that a balanced sample is simply not possible much of the time [12,14,22,31] due to cost, ethics, logistics, unavailability of cases, or simply not knowing what the sample will look like until the study is completed and the raters have generated their responses. The researcher should therefore review the relative proportion of cases in the categories and look for marked departures from a balanced sample. Observed agreement and specific agreement should be reported. If specific agreement is low in a rare category, then one simply cannot tell what the rater discrimination would have been with a more balanced sample.

In summary, with a severely unbalanced sample, measures like kappa and Finn's r are deceptive. Kappa appeared conservative in these examples, but when there are three categories, it is possible to have high kappa despite poor agreement on one of the categories if that category is rare. Specific agreement is the only measure that clearly points out the problem regardless of the number of categories. Specific agreement is not scaled like a reliability coefficient (chance agreement is not zero), however, and there is no way to tell from the sample what agreement would have been if the sample had been balanced.

6. Hypothesis testing

Depending on the size of a study sample, an estimate of the agreement metrics will differ from their underlying values. Testing the hypothesis that a metric differs from zero or from chance agreement is not generally useful, although such p-values have been reported in the medical informatics literature [32]. The fact that two raters agree by more than chance is not an appropriate measure for performance [33] and offers little comfort in creating a reference standard [15]. Raters can have poor discrimination but still agree by more than chance.

Only in a study where one is truly interested in the hypothesis that two raters differ by more than chance would a simple test of whether agreement differs from zero be appropriate [15]. This would not normally be the case in a reliability study, but kappa might be used to test for correlation among subjects in other studies. Davies and Fleiss [15] describes such an example of an epidemiological survey.

A more appropriate method for reliability studies is to calculate a 95% confidence interval for agreement and base one's conclusions on the lower bound of the interval [15]. To calculate a 95% confidence interval, one needs an estimate of the variance of the metric. For example, several approximations of the variance of kappa have been reported in the literature [9,10,15,34]. It is important to choose an approximation that gives the *non-null* standard error to calculate the confidence interval and to test whether kappa differs from a nonzero threshold. Many of the published formulae give the null standard errors and therefore are appropriate only to test whether the metric differs from zero.

Simple formulae for calculating the variances of agreement metrics are often not available, especially when there are more than two raters. Davies and Fleiss [15] and Dunn [9] recommend the use of computationally intensive methods such as the bootstrap or the jackknife [35]. Such techniques involve resampling the data to get an estimate of the variance of the empirical distribution. Normally one would resample on cases, but Dunn points out that one might want to resample on both cases and raters [9] if one wants to draw conclusions about raters in general rather than about the specific raters in the study. These methods can be used regardless of the number of raters or categories. The bootstrap estimate of the standard error of kappa for the example in Table 2 is .12, and the 95% confidence interval is .21 to .68.

7. Recommendation

7.1. General approach

Choosing how to quantify agreement is not simple, and the literature does not give a clear direction. Observed agreement is the most basic, easily understood metric, but it does not correspond to a reliability coefficient. It is nevertheless useful as an initial descriptive statistic to summarize the sample.

Kappa is perhaps the most popular metric, but the meaning of its magnitude (between 0 and 1) has probably been overinterpreted in experiments. Although most statisticians still consider it to have a broad and important role in quantifying agreement, some statisticians recommend that it essentially never be used, except in the limited circumstance of testing the null hypothesis of chance agreement with nominal data and more than two categories [36]. Its correspondence to an intraclass correlation coefficient and the ease of calculating it make it likely that it will remain popular.

The best approach depends on the goal. For example, if one is attempting to improve the reliability of some instrument, then it is important to pick an approach that separates the components of agreement. Specific agreement identifies disagreement on each category so that training can be targeted to a specific problem area. The marginal totals for the contingency table can reveal differences in rater thresholds (see Table 4, for example), and tests like McNemar's test and Cochran's Q [9] can determine the significance of the differences. Tetrachoric correlation, polychoric correlation, and latent trait models will also separate a difference in threshold from other sources of disagreement. The threshold component can then be improved through feedback and training.

If there is good reason to believe that a particular decision making model is operative, then that model should be employed in the analysis, as it may reveal more information about the disagreement [29]. Some aspects of the models can be tested empirically, but this does not justify a random search of approaches for one that appears to fit (and happens to report good agreement). Failure to reject a model does not prove that it is appropriate.

Unbalanced samples are particularly challenging, as it is impossible to accurately determine what reliability would have been in a more balanced sample. Specific agreement for each category is the most useful measure because it at least highlights where the difficulties lie.

7.2. Specific recommendations for nominal data (with at least three categories)

If the categories are nominal, then observed agreement and specific agreement on each category should be reported as descriptive agreement measures (but not as formal reliability measures). The prevalence of each category should be reviewed. Only if the sample is relatively balanced will an accurate assessment of reliability be possible. Kappa is the most common measure of reliability, but its level should not be overinterpreted. Like observed agreement and specific agreement, kappa should be compared only among studies of similar design, similar categories, and similar prevalence of observed values. A latent class model can be used if there are more than two raters, the model appears reasonable, and the model fits empirically.

7.3. Specific recommendations for ordinal data

If the categories are ordinal, and if one can assign some reasonable numeric score to each category, then it is probably best to use a method appropriate for continuous data such as a correlation coefficient [21] (e.g., intraclass correlation [18], product moment correlation [37]). For example, if the semantic intervals between adjacent categories are equal or if there is some concrete numeric interpretation of each category (e.g., a probability), then it may be best to treat the data as numeric. A common alternative for ordinal data is to use weighted kappa, but it requires setting relative weights for each type of disagreement, which is no less arbitrary than assigning numeric scores to each category. If the categories are ordinal, and if it is not reasonable to assign a numeric score to each one, then a model such as polychoric correlation or latent trait model may be reasonable.

Weighted kappa can be useful for categorical data that are not strictly ordinal but rather follow some more complex hierarchy [9]. Weights for disagreement between pairs of categories must be defined, and the magnitude of kappa must only be compared to that of similar experiments.

7.4. Specific recommendations for dichotomous data

If the categories are dichotomous, then the data can be treated as outlined above for nominal data. The methods for ordinal data can also be applied. Given two raters and two categories, however, there is only so much one can learn from the data, and models cannot be tested. Observed agreement, specific agreement, kappa, and tetrachoric correlation can be reported, but showing the two-by-two contingency table with its marginal totals is probably as informative as any measure.

Acknowledgments

Supported by National Library of Medicine grants R01 LM06910 “Discovering and applying knowledge in clinical databases” and R01 LM06274 “Unlocking data from medical records with text processing,” and Pfizer grant NY01-002153889 “Using information systems to advance clinical research and clinical care.”

References

[1] Friedman CP, Wyatt JC. Evaluation methods in medical informatics. New York: Springer; 1997.

[2] van der Lei J, Musen M, van der Does E, Man in't Veld AJ, van Bemmel JH. Comparison of computer-aided and human review of general practioners' management of hypertension. Lancet 1991;338:1504–8.

[3] Giuse NB, Giuse DA, Miller RA, et al. Evaluating consensus among physicians in medical knowledge base construction. Methods Inf Med 1993;32:137–45.

[4] Reggia JA, Tabb DR, Price TR, Banko M, Hebel R. Computer-aided assessment of transient ischemic attacks. Arch Neurol 1984;41:1248–54.

[5] Brown PJB, Sonksen PA. Evaluation of the quality of information retrieval of clinical findings from a computerized patient database using a semantic terminological model. J Am Med Inform Assoc 2000;7:392–403.

[6] Lenert LA, Tovar M. Automated linkage of free-text descriptions of patients with a practice guideline. Proc Annu Symp Comput Appl Med Care 1993:274–8.

[7] Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med 1995;122:681–8.

[8] Hripcsak G, Wilcox A. Reference standards, judges, comparison subjects: roles for experts in evaluating system performance. J Am Med Inform Assoc 2002;9:1–15.

[9] Dunn G. In: Design and analysis of reliability studies. New York: Oxford University Press; 1989. p. 154.

[10] Fleiss JL. In: Statistical methods for rates and proportions. 2nd ed. New York: Wiley; 1981. p. 212–36.

[11] Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait. Biometrics 1975;31:651–9.

[12] Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993:423–9.

[13] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.

[14] Cicchetti DV, Feinstein AR. High agreement but low kappa: II resolving the paradoxes. J Clin Epidemiol 1990;43:551–8.

[15] Davies M, Fleiss JL. Measuring agreement for multinomial data. Biometrics 1982;38:1047–51.

[16] Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213–20.

[17] Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76:378–82.

[18] Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.

[19] Koch JR, Landis GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.

[20] Carletta J. Assessing agreement on classification tasks: the kappa statistic. Comput Linguistics 1996;22:249–54.

[21] Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 1987;126:161–9.

[22] Feinstein AR, Cicchetti DV. High agreement but low kappa: I. the problems of two paradoxes. J Clin Epidemiol 1990;43:543–9.

[23] Thompson WD, Walter SD. A reappraisal of the kappa coefficient. J Clin Epidemiol 1988;41:949–58.

[24] Hripcsak G, Kuperman GJ, Friedman C, Heitjan DF. A reliability study for evaluating information extraction from radiology reports. J Am Med Inform Assoc 1999;6:143–50.

[25] StataCorp. Stata Statistical Software: release 4.0, vol. 2. College Station, TX: StataCorp; 1995. p. 163.

[26] Hutchinson TP. Kappa muddles together two sources of disagreement: tetrachoric correlation is preferable. Res Nurs Health 1993;16:313–5.

[27] Ubersax JS. Modeling approaches for the analysis of observer agreement. Invest Radiol 1992;27:738–42.

[28] Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 1974;61:215–31.

[29] Espeland MA, Handelman SL. Using latent class models to characterize and assess relative error in discrete measurements. Biometrics 1989;45:587–99.

[30] Pollack I, Norman DA. A non-parametric analysis of recognition experiments. Psychon Sci 1964;1:125–6.

[31] Whitehurst GJ. Interrater agreement for journal manuscript reviews. Am Psychologist 1984;39:22–8.

[32] Reggia JA. Evaluation of medical expert systems: a case study in performance assessment. Proc Annu Symp Comput Appl Med Care 1985:287–91.

[33] Hilden J, Habbema JD. Evaluation of clinical decision aids—more to think about. Med Inform (Lond) 1990;15:275–84.

[34] Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. Psychol Bull 1969;72:323–7.

[35] Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993.

[36] Ubersax JS. Available from: http://ourworld.compuserve.com/homepages/jsuebersax/agree/htm.

[37] Snedecor GW, Cochran WG. In: Statistical Methods. eighth ed. Ames: Iowa State University Press; 1989. p. 177.