RESEARCH METHODS & REPORTING

Clinicians are right not to like Cohen's κ

Henrica C W de Vet *professor of clinimetrics*¹, Lidwine B Mookink *junior researcher of clinimetrics*¹, Caroline B Terwee *assistant professor of clinimetrics*¹, Otto S Hoekstra *professor of nuclear medicine*², Dirk L Knol *assistant professor of statistics*¹

¹Department of Epidemiology and Biostatistics, EMGO Institute for Health and Care Research, VU University Medical Center, Amsterdam, Netherlands;

²Department of Radiology and Nuclear Medicine, VU University Medical Center, Amsterdam, Netherlands

Abstract

Clinicians are interested in observer variation in terms of the probability of other raters (interobserver) or themselves (intraobserver) obtaining the same answer. Cohen's κ is commonly used in the medical literature to express such agreement in categorical outcomes. The value of Cohen's κ , however, is not sufficiently informative because it is a relative measure, while the clinician's question of observer variation calls for an absolute measure. Using an example in which the observed agreement and κ lead to different conclusions, we illustrate that percentage agreement is an absolute measure (a measure of agreement) and that κ is a relative measure (a measure of reliability).

For the data to be useful for clinicians, measures of agreement should be used. The proportion of specific agreement, expressing the agreement separately for the positive and the negative ratings, is the most appropriate measure for conveying the relevant information in a 2x2 table and is most informative for clinicians.

Introduction

Observer variation is the Achilles' heel in clinical diagnoses such as medical imaging.¹ It directly affects the value of diagnostic tests and other measurements in clinical practice. Clinicians interested in observer variation pose questions such as "Is my diagnosis in agreement with that of my colleagues?" (interobserver) and "Would I obtain the same result if I repeated the assessment?" (intraobserver). To express the level of intraobserver or interobserver agreement, many epidemiology and medical statistics textbooks^{2,3} recommend Cohen's κ as the most adequate measure. Cohen introduced κ as a coefficient of agreement for categorical outcomes.⁴ Clinicians and researchers, however, have long been unhappy with κ because a high level of agreement can be accompanied by a low κ value. These counterintuitive results have led to dissatisfaction with κ expressed in papers with titles containing terms such as "paradox" and "bias" and calling for extensions and adjustments of κ .^{5,6}

An example of a confusing situation is provided by Bruynesteyn and colleagues in a study on rheumatoid arthritis that examined progression of joint damage based on pairs of radiographs from

46 patients.⁷ The authors compared the interobserver agreement between rheumatologists in two situations: firstly, when the pairs of radiographs were assessed in a chronological order and, secondly, when the rheumatologists had no knowledge about the order (random order). Table 1[↓] shows the proportion of observed agreement and the value of κ for these two situations. The highest value for the proportion of observed agreement was seen in the "chronological order" set, yet the highest κ value was seen in the "random order" set. The conclusion about which situation is preferred is therefore unclear.

In this paper we will first show how κ is calculated and explain these seemingly contradictory results. We will show how κ is a relative measure of observer variation, whereas the questions posed by clinicians require an absolute measure of agreement. Finally, we will focus on absolute measures of agreement.

Calculation of Cohen's κ

In their paper, Bruynesteyn and colleagues⁷ did not provide the raw data for the 2x2 tables. This gives us the opportunity to come up with numbers that mimic their results but are more attractive from an educational viewpoint. The first example presented by Bruynesteyn and colleagues might have resembled the data in table 2[↓].

The two rheumatologists, A and B, agree with each other on 38 out of the 46 sets of radiographs: in 33 cases they both observe progression and in five cases they both observe no progression. Therefore, the proportion of observed agreement (P_o) is $(a+d)/N=(33+5)/46=0.826$. Part of the observed agreement between two raters, however, is attributable to chance. As with multiple choice questions in an exam, some questions are answered correctly simply by guessing. Thus, in some cases, rheumatologist B might have agreed with rheumatologist A just by chance, even if neither of them had looked carefully at the radiographs. This is called "agreement by chance" or "expected agreement" (P_e). Cohen's κ adjusts for this expected agreement.⁴ Multiplying the row and column totals corresponding to each cell and dividing by the grand total (N) provides the expected

agreement for each cell in case of independent judgments. Therefore, the expected agreement in cell *a* is $(37 \times 37)/46 = 29.76$ cases, while the expected agreement in cell *d* is $(9 \times 9)/46 = 1.76$ cases. Consequently, the total proportion of expected agreement (P_e) amounts to $(29.76 + 1.76)/46 = 0.685$.

The formula for Cohen's κ is $(P_o - P_e)/(1 - P_e)$. In the numerator, the expected agreement is subtracted from the observed agreement. The denominator is also adjusted for the expected agreement. In this example, $P_o = 0.826$ and $P_e = 0.685$. Filling in the formula yields $\kappa = (0.826 - 0.685)/(1 - 0.685) = 0.45$.

Example 2 mimics the situation in which the two rheumatologists have examined the 46 radiographs in "random order." The results of the 2x2 table might have resembled the data in table 3.

In the second example 2 (table 3), we once again observe an agreement of 0.826 $((13 + 25)/46)$. The expected agreement is now $(18 \times 16)/46 = 6.26$ for "progression" and $(28 \times 30)/46 = 18.26$ for "no progression." This amounts to a proportion of 0.533 $((6.26 + 18.26)/46)$ for expected agreement. The resulting κ value is $(0.826 - 0.533)/(1 - 0.533) = 0.63$.

Example 2 shows the same proportion of observed agreement as example 1 (38/46 in both cases), but Cohen's κ value is higher. This is because of the difference in expected agreement. The row and column totals of the 2x2 table can be seen as the prevalence of progression, as observed by the two rheumatologists. The observed prevalence of progression was 80% (37/46) in example 1 and only 37% (17/46) in example 2 (averaged over two raters). The expected agreement is particularly high when the prevalence of abnormalities is either very high or very low (the prevalence of progression is 80% in example 1). It is in these situations that a high proportion of agreement can result in a low value of κ .

These examples show that a lower κ value is caused by higher expected agreement (as in example 1). It is still unclear, however, whether researchers should rely on the proportion of observed agreement or the κ value. To answer this question, we will first elaborate on relative and absolute measures of agreement.

Absolute and relative measures to quantify observer variation

Agreement represents an absolute measure and reliability is a relative measure (box 1 shows the formula and statistical details).^{8,9} The difference between agreement and reliability is most easily explained for situations of continuous outcomes, such as body weight. When a person's body weight is measured by different raters, we are interested in how much the observed weights differ. This is the absolute measure of variation—called the measurement error or extent of agreement—and can be expressed in units of measurement, such as kilograms. We want the measurement error to be as small as possible. Reliability is a relative measure as it relates the measurement error to the variation within a study sample. If the measurement error is small compared with the variation within the sample, the reliability is high. In a sample of adults, body weight varies from about 50 kg to more than 100 kg. This variation is much larger than that seen in birth weight, which might range from about 1.5 to 5 kg. Thus, babies form a more homogeneous sample than adults and therefore a measurement error of 0.5 kg in measurement of birth weight will lead to lower reliability. Consequently, weighing scales with this degree of error would not be suitable to measure birth weight.

Categorical outcomes, on the other hand, are based only on classifications and have no units of measurement. In these situations, closeness of the scores and measurement error are expressed as the probability of misclassifications or the proportion of observed agreement. The relative measure for categorical variables is Cohen's κ .

Cohen's κ is a reliability measure

Cohen's κ is a relative measure⁹ because it relates the proportion of observed agreement (the absolute measure) to variation in the sample and corresponds with a specific type of reliability measure.¹⁰ At this point, it is important to explain how variation is assessed in a sample with categorical or even dichotomous variables. In case of categorical outcomes, the sample is homogeneous if all patients have the same outcome (that is, are in the same category). A sample is maximally heterogeneous if the patients are equally distributed over the existing categories (that is, in case of dichotomous outcomes a 50-50 distribution). So, the row and column totals of the 2x2 table—that is, the prevalence of abnormalities or normalities—give an indication of the heterogeneity of the sample. In case of a prevalence of 50% (50-50 distribution), the expected agreement will be minimal. With a higher or lower prevalence—that is, a more skewed distribution—the expected agreement will be greater and, assuming the same proportion of observed agreement, the value of κ will be smaller. This explains the findings in examples 1 and 2 (tables 2 and 3): while the absolute agreement is the same, the value of κ (as a relative measure) is lower in example 1 because the distribution of progression in the sample is more skewed.

So, whether we should rely on Cohen's κ or on a measure of agreement now comes down to whether the clinical question concerns reliability or agreement.

Do clinical questions concern reliability or agreement?

In clinical practice, clinicians perform assessments to diagnose and monitor individual patients. They pose questions regarding observer variation such as "Will my diagnosis be in agreement with a colleague's diagnosis?" and "Can I distinguish patients with abnormal scores from those with normal scores?" The first question clearly concerns agreement and requires an absolute measure of agreement. Thus, in a situation of continuous outcomes, such as body weight, the clinicians would want to know how close the observed weights were in terms of absolute difference in the units of measurement, and, in case of categorical variables such as radiological classifications, they would want to know the probability of agreement.

When clinicians ask whether patients with abnormal scores can be distinguished from those with normal scores, they also have individual patients in mind and are not considering the distribution of a sample of patients. This is illustrated by the radiological assessment of progression in patients with rheumatoid arthritis (tables 2 and 3). We see that the prevalence of progression is higher when rheumatologists judge the radiographs in chronological order rather than in random order (80% v 37%). They more often observe "progression," which means that they tend to label smaller differences as progression in the chronological order set. The observed agreement between the two observers is the same, but the relative measure is lower because the distribution is more skewed. Thus, taking the sample distribution into account results in a relative measure, which is not what clinicians have in mind when they ask whether they

Box 1: Formula and interpretation of a reliability measure

The formula for a reliability measure (Rel) of continuous outcomes is:

$$\text{Rel} = (\sigma_p^2) / (\sigma_p^2 + \sigma_{\text{error}}^2)$$

A reliability measure relates the measurement error variance (σ_{error}^2) to the variability between people (σ_p^2). In this formula, σ_{error}^2 represents how close the scores of repeated assessments are. The square root of σ_{error}^2 equals the standard error of measurement, which is expressed in the units of measurement—for example, 0.5 kg in case of weighing scales. When measuring body weight in adults, σ_{error}^2 will be small compared with the variation in the sample (σ_p^2), and the reliability measure will be close to 1 (excellent reliability), but when σ_{error}^2 is large compared with the variation in the sample (for example, when measuring birth weight) the reliability measure will be closer to 0 (poor reliability).

can distinguish between people with abnormal and normal scores.

Reliability is at stake if we want to know whether a specific test is suitable for identifying or classifying patients in a certain sample. For example, in the case of screening the underlying question is: “Can we identify patients with abnormalities in a large population sample?” Another reliability question concerns the evaluation of a new classification system for a test or measurement instrument: one might question whether raters can distinguish between the proposed categories. These issues involve the ability to distinguish between patients in population samples. Thus, when variation within the sample is of interest, reliability measures are preferred. This is the case when the research question refers to suitability of a test in a particular setting. In clinical practice, clinicians are interested in decision making in individual patients and therefore absolute measures are of interest.

Proportion of specific agreement as preferred measure of agreement

If rheumatologists have to decide which measure they would prefer to use to rate progression, they should consider an absolute measure of agreement as κ fails to provide any information that would help them interpret the results for their patients in clinical practice.

One such measure could be the proportion of observed agreement. The question, however, is whether this measure is sufficiently informative. The best information is provided by the complete 2x2 table itself. In table 2, the problem of misclassification seems to be most marked in the case of “no progression,” where the numbers on which the rheumatologists disagree (cells *b* and *c*) outweigh the numbers on which they do agree (cell *d*). On the other hand, there is no problem with the classification of “progression” (cell *a* versus cells *b* and *c*). This dual information cannot be captured in a single statistic and therefore the proportion of observed agreement is not sufficiently informative either. The measure that would convey this information best is the proportion of specific agreement. This is a measure that expresses the agreement separately for positive and negative ratings—that is, agreement on the diagnosis of “progression” and on the diagnosis of “no progression.” The specific agreement on a positive rating, known as the positive agreement, is calculated by the following formula¹¹: $PA = 2a / (2a + b + c)$, while specific agreement on a negative rating, the negative agreement, is calculated using the formula: $NA = 2d / (2d + b + c)$. The inclusion of both cells *b* and *c* in the formula accounts for the fact that these numbers might be different (see table 3), and therefore their mean value is taken.

The proportion of specific agreement helps the clinician answer questions such as: “Suppose I rate ‘progression’ based on a pair of radiographs, what is the probability that another clinician would also rate ‘progression’?” In the case of example 1 (table 2), where the observed prevalence of progression was 80%, the answer would be $PA = (2 \times 33) / ((2 \times 33) + 4 + 4) = 0.892$ —that is,

89.2% for agreement on “progression”—and $NA = (2 \times 5) / ((2 \times 5) + 4 + 4) = 0.556$ —that is, 55.6% for agreement on “no progression.”

In the case of example 2 (table 3), where the prevalence of progression was 37%, the probability that other clinicians would also observe progression would be $PA = (2 \times 13) / ((2 \times 13) + 3 + 5) = 0.765$ —that is, 76.5% for “progression”—and $NA = (2 \times 25) / ((2 \times 25) + 3 + 5) = 0.862$ —that is, 86.2% for “no progression.”

Note that in the above formulas the same numbers of disagreements (that is, cells *b* and *c*) are related to the agreements on the positive ratings (cell *a*) and to the agreements on the negative ratings (cell *d*). We see again the influence of prevalence, but now it has direct clinical meaning as it is transformed into a probability of agreement for positive and negative ratings. When a sample is homogeneous because almost all patients show progression (example 1), the probability of observer agreement on progression is higher. Similarly, when only a small number of patients have no progression, the probability of observer agreement on non-progression becomes smaller, despite the same numbers of misclassification. Consequently, the proportion of specific agreement corresponds better with the reasoning of a clinician. Moreover, distinguishing between the proportion of agreement for positive and negative ratings makes sense because both outcomes have different clinical consequences.

It is interesting to note that the measure of specific agreement was first described by Dice in 1945¹¹ and later revitalised in the medical literature by Cicchetti and Feinstein.¹² Until now, however, it has not found broad application, despite being an extremely helpful measure for clinicians.

Conclusion

As a mnemonic, clinicians might like to remember that Agreement is an Absolute measure and that RELiability is a RELative measure to quantify observer variation. Cohen introduced κ as a measure of agreement and many authors have followed suit as the need to take chance agreement into account sounds quite plausible. Adjustment of the observed agreement for the expected agreement, however, turns κ into a relative measure. Therefore, we should stop adapting Cohen’s κ and instead be critical about whether a specific clinical question asks for a measure of reliability or a measure of agreement. In particular, Cohen’s κ is not to be recommended as a measure of observer variation in clinical practice. Such questions regarding a clinician’s confidence in a specific diagnosis concern agreement and, as such, require an absolute measure of agreement. The measure of specific agreement best conveys the relevant information contained in a 2x2 table and is most helpful for clinical practice.

Contributors: All authors had a substantial contribution to the conception and design of the paper. HCWdV drafted the paper, and all other authors

Summary points

- As Cohen's κ was originally introduced as a coefficient of agreement for categorical variables, researchers often use Cohen's κ to express observer agreement in clinical diagnoses
- Cohen's κ is falsely known as an agreement measure, whereas it is a measure of reliability. Agreement measures are absolute measures, and reliability measures are relative measures relating the absolute agreement or measurement error of a characteristic to its variation in the sample
- The proportion of specific agreement, rather than Cohen's κ , is an informative agreement measure for clinicians

added intellectual content and critically reviewed and revised the paper. All authors have approved the final version. HCWdV is guarantor.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Provenance and peer review: Not commissioned; externally peer reviewed.

- 1 Robinson PJ. Radiology's Achilles' heel: error and variation in the interpretation of the Röntgen image. *Br J Radiol* 1997;70:1085-98.
- 2 Sackett D, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology—a basic science for clinical medicine. Little, Brown and Company, 1991:25-31.
- 3 Altman DG. Practical statistics for medical research. Chapman and Hall/CRC Press, 1991:403-9.
- 4 Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960;20:37-46.

- 5 Feinstein AR, Cicchetti DV. High agreement but low kappa: I. the problems of two paradoxes. *J Clin Epidemiol* 1990;43:543-9.
- 6 Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol* 1993;46:423-9.
- 7 Bruynesteyn K, Van Der Heijde D, Boers M, Saudan A, Peloso P, Paulus H, et al. Detecting radiological changes in rheumatoid arthritis that are considered important by clinical experts: influence of reading with or without known sequence. *J Rheumatol* 2002;29:2306-12.
- 8 De Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol* 2006;59:1033-9.
- 9 Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hrobjartsson A, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol* 2011;64:96-106.
- 10 Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 1973;33:613-9.
- 11 Dice LR. Measures of the amount of ecologic association between species. *Ecology* 1945;26:297-302.
- 12 Cicchetti DV, Feinstein AR. High agreement but low kappa: II Resolving the paradoxes. *J Clin Epidemiol* 1990;43:551-8.

Accepted: 04 March 2013

Cite this as: *BMJ* 2013;346:f2125

© BMJ Publishing Group Ltd 2013

Tables

Table 1 | Observed agreement and κ values for “chronological order” and “random order” sets (adapted from Bruynesteyn et al⁷)

	Prevalence of progression of joint damage	Observed agreement (proportion)	Cohen's κ
Chronological order	80%	0.85	0.45
Random order	37%	0.80	0.60

Table 2| Example 1: 2x2 table of “chronological order” dataset

Rheumatologist B	Rheumatologist A		Total
	Progression	No progression	
Progression	33 (a)	4 (b)	37
No progression	4 (c)	5 (d)	9
Total	37	9	46 (N)

Table 3| Example 2: 2x2 table of “random order” dataset

Rheumatologist B	Rheumatologist A		Total
	Progression	No progression	
Progression	13 (a)	3 (b)	16
No progression	5 (c)	25 (d)	30
Total	18	28	46 (N)