Interventional Pain Medicine 1 (2022) 100124

Contents lists available at ScienceDirect

Interventional Pain Medicine

journal homepage: www.journals.elsevier.com/interventional-pain-medicine

On understanding reliability for diagnostic tests

Nikolai Bogduk

The University of Newcastle, PO Box 431, East Maitland, NSW, 2323, Australia

ARTICLE INFO

Keywords:
Reliability
Agreement
Diagnostic test
Kappa

ABSTRACT

For professional practice to be responsible, any diagnostic tests used must be reliable. Therefore, the reliability of any diagnostic test needs to have been measured. The classical statistic for quantifying reliability is Kappa. Although Kappa can be promptly determined using a programmed calculator, using an algorithm to derive Kappa provides greater insight into what it is actually measuring and why. Kappa scores can be graded, with verbal descriptor applied to different grades. However, those descriptors do not necessarily reflect the degree of skill required to achieve different grades of Kappa. High levels of skill attract high Kappa scores, but Kappa scores described as fair or moderate are not necessarily flattering because they can be achieved with questionable levels of skill. Various corrections can be applied to the calculation of Kappa scores in order to raise their value, and to improve the verbal descriptors of their grade, but these may not be legitimate or necessary. Low Kappa scores do not condemn tests but they serve to raise questions about their reliability.

1. Introduction

Some physicians have difficulty grasping the concept of reliability and its significance in the evaluation of diagnostic tests. Intuitively, they are comfortable with the concept validity, meaning if the result of the test is correct or not; but reliability seems more elusive or irrelevant.

Whereas validity deals with the results of the test, reliability pertains more to how well the test is performed, and how its results are interpreted. Therein, reliability deals with the human elements involved in conducting the test. A physician might obtain a medical image of a patient, but someone has to read the film and interpret what it shows. A physician might palpate a patient, but has to interpret what they have felt.

These human factors can involve differences in what the physician sees, feels, detects, or thinks. These differences can lead to inconsistencies between physicians in deciding whether the result of the test is positive or negative.

Consistency is crucial to the clinical utility of any diagnostic test. If a test lacks consistency, one physician might find the test positive in a given patient, but any number of other physicians might find the test negative, were they to examine the same patient. Under such conditions, the test is useless for making a decision, because there is no way of determining which physician is correct. In order to resolve disputes, the diagnosis would need to be checked by some other, less flawed test; but if such a test is available, it should have been used in the first instance, in order to avoid generating doubt and disagreement.

2. Definitions

Reliability can be defined as the extent to which two (or more) observers, using the same diagnostic test on the same group of patients, agree on the results of the test. Since two or more observers are involved, this reliability is known as inter-observer reliability.

A special case can be formulated when only one observer is involved. The reliability of a single observer can be defined as the extent to which a single observer obtains the same results in each patient when they re-test the same group of patients. This is known as intra-observer reliability. It measures how consistent a single observer is in agreeing with themselves.

3. Measuring reliability

Logistically, measuring reliability is a straightforward exercise. It requires two observers each to apply the same test in the same group of patients, and record their results. Those results can then be entered into a contingency table (Table 1). Such a table shows the raw data distributed according to the numbers of patients in which the observers agreed or disagreed on the results of their tests.

In such a table, the cells “a” and “d” constitute what are qualitatively good results. Both observers agreed either that the test was positive or the test was negative. Conversely, cells “b” and “c” are bad results, because the observers did not agree.

If a diagnostic test is tested in this way, and if agreement is found to be poor, the study itself does not tell us the reason for the lack of agreement.

E-mail address: nbogduk@bigpond.net.au.

https://doi.org/10.1016/j.inpm.2022.100124
Received 13 June 2022; Accepted 16 June 2022

Table 1
The components and structure of a contingency table for measuring agreement. The numbers in the cells (a,b,c,d) are the numbers of patients for whom the observers recorded the result of the test respectively defined by the row and the column that intersect the cell.

		Observer 1	
		Test Positive	Test Negative
Observer 2	**Test Positive**	a	b
	Test Negative	c	d

Sometimes there might be an inherent flaw in the mechanisms of the test itself. More often the flaw lies in how the test is performed or interpreted by the user. Additional studies would be required to find the responsible reason. However, irrespective of the reason, testing the test in this way can reveal, in the first instance, that something is suspect about the test or its performance.

Conducting research into the reliability of diagnostic tests is not an academic indulgence or an arbitrary option. For maintenance of quality clinical practice, it is imperative that the reliability of any test be known. If reliability has not been measured, users have no way of knowing if the test works properly or not, or if they are performing it correctly. Assuming that it works, or relying on wishful thinking, is not a substitute for actual evidence.

4. Classical statistics

Introduced in 1960 by Cohen [1], the classical statistic for measuring reliability is Kappa, also known as Cohen's Kappa. For the generic data of Table 1, the value of Kappa can be expressed by the equation:

$$Kappa = \frac{[a+d]-[a+c][a+b]-[b+d][c+d]}{[a+b+c+d]-[a+c][a+b]-[b+d][c+d]}$$

Such an expression is both visually overwhelming and ambiguous. It provides no insight into exactly what Kappa is measuring or what it is indicating.

For practical purposes, physicians do not need to memorise this equation. For a given set of data, they can use any number of calculators for Kappa that are available on the Internet. However, readers who are prepared to be patient can follow an algorithm that not only derives the statistic but also explains the logic behind it. Furthermore, patient readers might discover that it is a lot easier to remember the logic of the algorithm than it is to remember a complex equation. Armed with that logic, they will always be able to calculate the statistic, with insight, and without resorting to the Internet.

The data in Table 1 show that the observers agreed in a total of [a] cases that the test was positive, and in [d] cases that the test was negative. This implies that their overall rate of agreement is [a+d] cases out of a total of [a+b + c + d] cases. However, this constitutes only the "apparent" agreement. Finding the "true" agreement requires adjusting the raw data for chance agreement.

The first step in finding the chance agreement is to add a column and a row to Table 1, in order to create what are called margins to the table. Into these margins we add the totals of the respective columns and rows, as shown in Table 2. Next we retain these totals but erase the original data inside the central cells (Table 3).

Table 2
A contingency table measuring agreement in which the sums of the columns and rows have been added to the marginal column and marginal row.

		Observer 1		
		Test Positive	Test Negative	
Observer 2	**Test Positive**	a	b	a + b
	Test Negative	c	d	c + d
		a + c	b + d	N = a+b + c + d

Table 3
A contingency table for measuring agreement in which the original data within the central cells have been erased but the marginal totals have been retained.

		Observer 1		
		Test Positive	Test Negative	
Observer 2	**Test Positive**			a + b
	Test Negative			c + d
		a + c	b + d	N = a+b + c + d

The marginal totals in Table 3 indicate what the "average" behaviour is of each observer. On average, Observer 1 rates [a + c] cases out of N as positive, and [b + d] cases out of N as negative. This tells us that if Observer 1 operates by chance alone (for example by guessing) they will rate [a + c]/N cases as positive regardless of how many cases are presented to them. Likewise, they will rate [b + d]/N cases as negative. For Observer 2, the respective rates are [a + b]/N and [c + d]/N. We can use this information to backfill the contingency table with numbers that would have arisen if the two observers were operating simply by chance.

Observer 1 rated [a + c] cases as positive. If, hypothetically, we were to present these [a + c] cases to Observer 2, he or she would, on average, rate [a + b]/N of these cases as positive. This provides us with a number [a_{chance}] that is the number of agreements for positive cases achieved by chance alone, and whose value is calculated as:

$$a_{chance} = [\frac{[a+b]}{N}].[a+c]$$

Observer 1 rated [b + d] cases as negative. If we were to present these cases to Observer 2, he or she would, on average rate [c + d]/N of these cases as negative. So, the number of agreements for negative cases achieved by chance alone would be:

$$d_{chance} = [\frac{[c+d]}{N}].[b+d]$$

We can now enter these two values into Table 3, as shown in Table 4. (We do not need to calculate values for the "b" and "c" cells because these cells measure disagreement, and are not taken into consideration when calculating agreement.

We now have at our disposal two numbers that we can start to compare. Table 1 provides us with [a + d]/N, which constitutes the observed rate of agreement (O) according to the raw data. Table 5 provides us with [a_{chance} + d_{chance}]/N, which is the rate of agreement that we would expect (E) to have occurred by chance alone. We can now compare these two rates to see the extent to which the two observers were operating beyond chance alone.

The logic behind this comparison is illustrated in Fig. 1. Fig. 1A depicts a rectangle that metaphorically depicts the total **range of possible agreement**, from 0 to 1, where 1 represents total or 100% agreement. Fig. 1B shows that some of range of possible agreement is occupied by the **expected agreement** by chance alone, which has a magnitude of [E]. This range is not a fixed proportion; it differs from study to study depending on the expected agreement calculated from the raw data for each study. Once the magnitude of the expected agreement [E] has been established the remainder of the total range of possible agreement becomes the range for **possible agreement beyond chance**, and has a

Table 4
A contingency table for measuring agreement in which the marginal totals have been used to calculate the values of the "a" and "d" that would be expected had the observers been operating by chance alone.

		Observer 1		
		Test Positive	Test Negative	
Observer 2	**Test Positive**	a_{chance}		a + b
	Test Negative		d_{chance}	c + d
		a + c	b + d	N = a+b + c + d

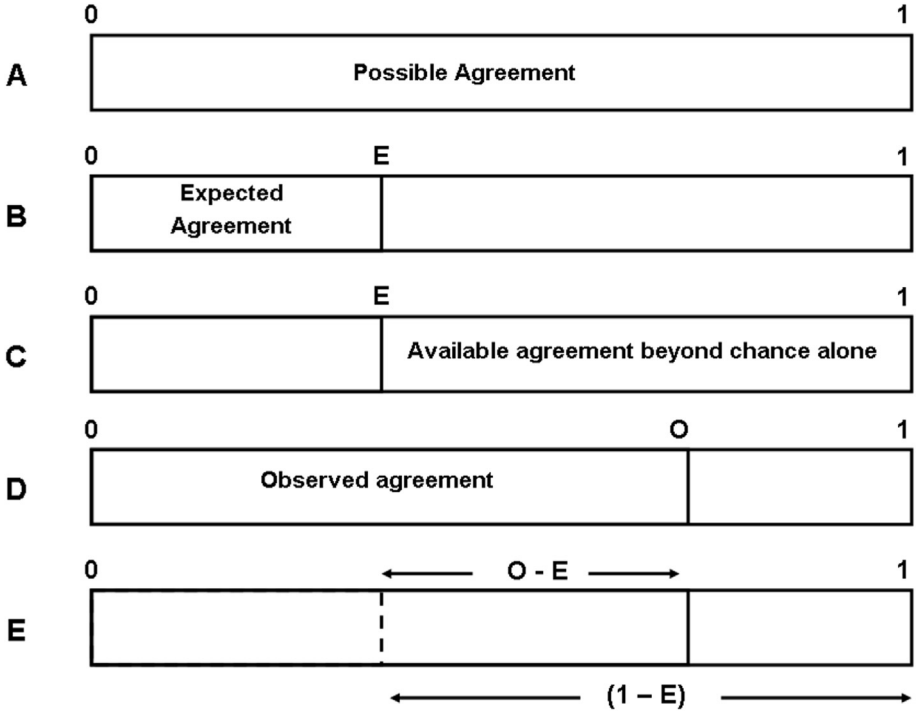

Fig. 1. A graphic representation of the derivation of a kappa score for agreement. **A:** For any test there is a range of possible agreement. **B:** Some of the possible agreement is taken up by expected agreement that occurs by chance alone. **C:** The remaining range of possible agreement is the available agreement beyond chance alone. **D:** The observed agreement encompasses the expected agreement and a residual that extends into the range of agreement beyond chance. **E:** True skill is represented by the extent to which the residual agreement (O–E) extends into the available range beyond chance alone (1-E).

magnitude of [1 – E] (Fig. 1C).

We can now examine how the **observed agreement** [O] compares with these two preceding measures. Typically, the observed agreement will be larger than the expected agreement, as shown in Fig. 1D, although not always so (see below). The observed agreement [O] consists of the expected agreement [E] plus a residual, whose magnitude is [O – E]. In essence, the observed agreement [O] is discounted by the expected agreement [E] to generate the **observed agreement beyond chance** [O – E] (Fig. 1E).

The definition of true agreement (Fig. 1E) is the extent to which the observed agreement beyond chance (O – E) extends into the range of available agreement beyond chance (1 –E), and is expressed as a proportion, i.e.

$$\text{True agreement} = \frac{O - E}{1 - E} = Kappa$$

where Kappa is a singular statistic that reflects the magnitude or the strength of the true agreement.

Calmly reflecting on this simple equation and Fig. 1 shows that:

1. If the observers have perfect agreement, the observed agreement [O] will extend across the entire range of [1 – E]. In mathematical terms, [O – E] will equal [1 – E], and Kappa will equal 1.0.
2. If the observers are simply guessing, they will show no skill beyond chance, which means that [O – E] does not extend at all into the range of [1 – E]. Mathematically, this is expressed as [O] equals [E]. So, [O – E] equals zero, and Kappa becomes 0. A Kappa score of zero reflects that the observers have no nett skill beyond chance alone.

When the observed agreement [O] is less than the expected agreement [E], the nett agreement [O – E] becomes less than 1, and Kappa assumes negative values, across a range from 0 to −1.0. Although somewhat unusual, such values have arisen in some studies of agreement. Negative scores indicate either that the quality of the test in question or its execution is so poor that the results generated are consistently worse than simply guessing.

The concept of discounting the observed Kappa score can be reinforced by a comparison. Let there be an examination, consisting of 100 multiple choice questions, each with four options but only one correct. A candidate answers 80 questions correctly. Is the candidate's level of skill 80%? The answer is no. There is a confounder in the design of the examination that affects how it can measure skill. Anyone with no skill could sit for the examination and, on average, score 1 in 4 questions correctly simply by answering randomly. Thereby, anyone – and everyone – can score 25% simply by chance. The true skill of the candidate is expressed by the extent to which they answer the remaining 75 questions correctly. So, a score of 80 questions correct is discounted by 25, leaving 55. The available number of questions beyond chance is 100–25 = 75. The candidate's true skill is 55/75 = 73%.

5. Interpretation

As explained in Appendix 1, and as illustrated in Fig. 2, Kappa scores are related to the skill of the observers performing the test but are not numerically equivalent. This arises because whereas levels of skill are anchored at zero and 100%, Kappa scores are anchored at skill levels between 50% and 100%. Kappa is zero for a skill level of 50% because that level of skill cannot be distinguished from random guessing.

In effect, Kappa rewards good levels of skill but is unrewarding for lower levels of skill. High levels of skill are accorded high-range Kappa scores, but Kappa draws attention to low levels of skill by according them low-range scores. Moderate levels of skill attract modest Kappa scores, which might seem low, but perhaps not unduly so.

One way of understanding the properties of Kappa is to appreciate that it is not designed to reward or celebrate high levels of agreement. Kappa serves more to identify, or to bring under suspicion, levels of agreement that suggest low levels of skill. So, Kappa is more like a screening test for poor agreement rather than for good agreement.

This effect is evident in Fig. 2. In zone A of Fig. 2, the numerical values of the Kappa scores (were they to be expressed as a percentage) are lower than the corresponding numerical values for the skills, but broadly speaking, high skills get high Kappa scores, and vice versa. In zone B, modest levels of skills are accorded low Kappa scores. This is not unreasonable. Skill levels of around 60% are perilously close the guessing level of 50%, and should not be accorded flattering Kappa scores.

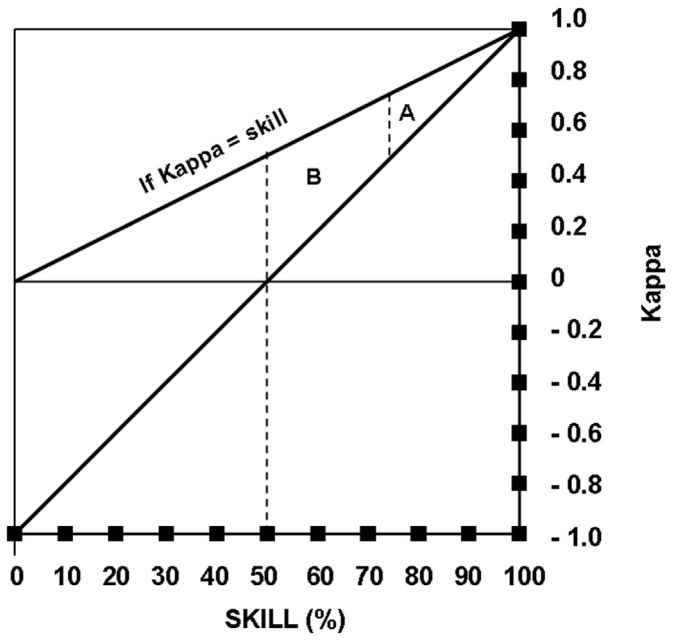

Fig. 2. A graph showing how Kappa scores are related to skill, and a hypothetical line showing the relationship if Kappa equalled skill directly. The graphs show that Kappa diverges from measuring skill directly. In Zone A, high levels of skill attract high Kappa scores, but the Kappa scores (if expressed as a percentage) understate the skill slightly. In Zone B, Kappa scores are substantially understate the associated level of skill. In effect, the divergence between Kappa and skill shows that Kappa penalises lower levels of skill, and makes them more obvious by according them low scores.

6. Translation

From time to time, various contributors to the literature have offered sets of adjectives used to translate the numerical value of Kappa scores into words (Table 5). These adjectives serve to convey comparative, qualitative values to the levels of agreement that the Kappa scores reflect.

However, the risk applies that investigators can use these same terms to indicate the quality of the diagnostic test that has been assessed. In some instances this might be reasonable. A test that produces strong or good agreement is likely to be of high or good quality. In other instances the translation may be inappropriate. When applied to the quality of the test, rather than to the agreement, descriptors such as moderate or fair can be unduly flattering to a test of questionable quality. Investigators intent on promoting a particular diagnostic test can be attracted to use such flattering interpretations to vindicate their test, despite what the Kappa score actually indicates.

Table 6 shows an alternative set of descriptors. The descriptors for agreement are objective, and relate simply to the numerical value of Kappa. Subjective descriptors are provided separately for the implied quality of the test. Very high Kappa scores are rarely achieved in studies of clinical tests. Therefore, tests that achieve such scores qualify for being recognised as of outstanding quality. High Kappa scores imply tests that, in practice, achieve agreement in more than 80% of cases, and which would be considered as acceptably good quality for most clinical purposes.

Conversely, low and very low Kappa scores indicate tests that more often than not fail to achieve agreement beyond chance. Such tests do not assist clinical practice because they generate doubt. Accordingly they are rated as poor quality.

The grey zone pertains to Kappa scores in the range between 0.4 and 0.6. Mathematically, these are medium range values but they do not indicate tests of medium quality. Tests with Kappa scores above 0.5 are on the verge of high agreement and good quality and so, might be considered as potentially good quality. They might be tolerable for some clinical situations, or their quality might be improved through refinement or better training. Tests with Kappa scores below 0.5 fail more often than not to achieve agreement in practice and, therefore, should be called into question. Other studies might produce more flattering Kappa scores for these tests, or the tests need to be refined or their performance improved.

Any set of verbal descriptors carries inherent personal judgement values. Readers are free to adapt Table 6 to reflect their own personal values. The guiding principle, however, would be to ask: just how good should a test be for responsible clinical practice. Intuitively, tests that reflect secure agreement in 80% or 90% of cases seem to achieve that threshold. Perhaps a skill level of 75% might be tolerable. It becomes a matter of conscience to decide if skill levels of 70% or 60% reflect good clinical practice.

Table 5
Suggested verbal descriptors for the agreement implied by various ranges of values of Kappa.

Kappa	Descriptor		
1.0			
	Almost perfect	Strong	Very good
0.8			
	Substantial	Moderate	Good
0.6			
	Moderate	Weak	Moderate
0.4			
	Fair	Minimal	Fair
0.2			
	Slight	None	Poor
0.0			
Source	Landis, Koch [2]	McHugh [3]	Altman cited by McHugh [3]

Table 6
Suggested verbal descriptors for the agreement indicated by various ranges of values of Kappa, and the respective implications for the quality of the test.

Kappa	Agreement	Test Quality
1.0		
	Very High	Outstanding
0.8		
	High	Good
0.6		Potential
	Medium	Questionable
0.4		
	Low	Poor
0.2		
	Very Low	Very Poor
0.0		

7. Confidence intervals

Because Kappa is calculated as a proportion it inherits the mathematical liabilities of proportions. Variations in the raw data used to calculate the value of Kappa can generate differences in that value. Different studies of the same diagnostic test may yield different values that estimate, but do not necessarily establish, the true value of Kappa for that test.

From any given set of data, the range in which the true value of Kappa lies can be calculated using the 95% confidence limits of the observed value. Those limits can be approximated by the equation [1,3]:

$$Limits = Kappa \pm 1.96\sqrt{\frac{[O][1-O]}{n[1-E]^2}}$$

Calculating the confidence intervals of Kappa is not simply an academic ritual. Doing so can provide insight into the quality of the study that reports a Kappa value.

Table 7 shows the data of a hypothetical study of agreement that produced a Kappa score of 0.48. The question that arises is if this is a

Table 7
A contingency table for measuring agreement from a hypothetical study that enrolled 54 subjects.

		Observer 1	
		Test Positive	Test Negative
Observer 2	**Test Positive**	20	8
	Test Negative	6	20
		Kappa = 0.48	

reasonable and generalisable value of Kappa, i.e. is it the value that a reader would expect to encounter if they adopted the test that had been studied?

The answer to that question is provided by calculating the 95% confidence intervals of this observed Kappa. Those values are 0.25–0.71. This tells the reader that the true value of Kappa is not 0.48, but could be any value in the range from 0.25 to 0.71. It also tells the reader that the conclusion of the study should not be "Kappa equals 0.48" but "Kappa could be as low as 0.25 or as high as 0.71". The latter conclusion indicates that the study is not informative. The true Kappa could be low, medium, or high; and the study does not help us discriminate. Using the confidence intervals in this way informs the reader that the study is flawed by having too small a sample size.

Table 8 overcomes this problem. It shows the data of a study of the same diagnostic test, but with a larger sample size, and with raw data that are in the same proportions as those in Table 8. The new data generate the same value of Kappa (0.48), but its confidence intervals are 0.40–0.56. The reader can, therefore, accept these data as informative. Were the reader to adopt the test in question, that can be 95% confident that, in their own practice, they will encounter a Kappa score within the range of 0.40–0.56, which is close enough to the reported value of 0.48.

8. Concerns

A number of authors have raised concerns about the suitability and accuracy of Kappa as a measure of reliability. Some of these pertain to the mathematics of Kappa. Others pertain to the expectations and judgements of investigators.

A mathematical feature of Kappa is that, for any given sample size, the numbers in a contingency table operate in a closed system. There is a finite number of possible permutations of the numbers within the cells. As values in individual cells change, the ratios between cells change; but those changes are not linear; they depict asymptomotic curves. One consequence of this behaviour is paradoxical behaviour of Kappa [4,5]. As the prevalence of the index condition in the sample studied decreases or increases the value of Kappa decreases. Consequently, studies are disadvantaged, by achieving an unflattering Kappa, if the raw data are unbalanced, because they have too many or too few subjects with the condition being diagnosed. The implication of this peculiarity is that studies should be designed to have roughly an equal number of subjects with and without the condition. Thereby, observers have an equal opportunity both to find positive cases and correctly find negative cases. This provides for an optimal measure of Kappa.

A similar problem arises if and when the observers in a study perform differently [4,5]. This is referred to as observer bias. When the bias is large, the raw data are unbalanced, and Kappa appears to be understated.

Table 8
A contingency table for measuring agreement from a hypothetical study that enrolled 432 subjects.

		Observer 1	
		Test Positive	Test Negative
Observer 2	**Test Positive**	160	64
	Test Negative	48	160
		Kappa = 0.48	

For samples where the prevalence has not been 50%, some authorities recommend adjusting for prevalence, using a prevalence index, to produce a Prevalence-Adjusted Kappa [6,7]. The prevalence index essentially averages out the prevalence, and is applied a correcting coefficient in the calculation of Kappa.

A similar process can be used when a study finds that the observers do not behave in the same way. Based on the difference between the numbers of disagreements between observers, a bias index can be derived, and applied to calculate a Bias-Adjusted Kappa [6,7]. This adjustment can be added to prevalence adjusted to produce a composite Prevalence-Adjusted-Bias-Adjusted Kappa (PABAK).

Although these adjustments succeed in improving Kappa scores, readers are entitled to ask: are they necessary, and are they legitimate. Without adjustments, Kappa already recognises tests of good quality by according them high scores. It is only in the context of poor or moderate scores that the need for adjustments appears to arise; but low, unadjusted Kappa scores do not condemn a test, they only raise suspicion about them. Making adjustments to the results of a poorly designed study can seem to be like camouflaging the truth. Adjustments might serve as an indication of what the true Kappa score could be, but they are not a substitute for a better conducted study. If prevalence is an issue, a balanced sample should be used instead of a prevalence index. If observer bias is an issue, investigating the reason for that bias sounds more responsible than averaging out the raw data with a bias index.

Gwet's agreement coefficient [8,9] is relatively new statistic that has evolved from Kappa. Although akin to Kappa in broad terms, this coefficient is calculated in a mathematically more complex manner, to reduce the penalty for chance agreement. It was developed on the grounds that Kappa is unduly punitive for chance agreement, and in order to serve the assumption that skilled observers would be unlikely to be benefitting from chance agreement and so, should not be penalised. Comparison studies have shown that Gwet's agreement coefficient generates more flattering scores than does Kappa. However, the assumption that observers are not likely to be relying on chance agreement can be questioned. Although diplomatic and complimentary, this assumption has not been shown to be valid for all tests in medical practice, and particularly not for those tests related to pain medicine. For those tests, a low Kappa score might be a closer reflection of reality than a score that lowers the discount for guessing.

9. Studies

Many factors can affect the quality of studies of reliability, other than the statistics used to measure it. Table 9 summarises these factors, and serves as a guide to readers as to what to look for if and when they need encounter studies of reliability. The list is also a guide to planning a study, so as to avoid predictable flaws. A detailed explanation and discussion of these factors is available in the literature [10,11].

10. Diagnostic blocks

Diagnostic blocks have been given a rationale and guidelines for their execution to ensure their validity [12,13]. However their reliability has

Table 9
A list of considerations when reading or planning a study of reliability.

Was the sample of subjects representative?
Was the sample of raters representative?
Were raters blinded to the findings of other raters?
Were raters blinded to their own prior findings?
Were raters blinded to the accepted reference standard?
Were raters blinded to clinical information not part of test?
Were raters blinded to additional non-clinical cues?
Was the order of examination varied?
Was the time interval between repeated measures appropriate?
Was the test applied correctly and interpreted appropriately?
Were appropriate statistical measures of agreement used?

not been investigated. It has been convenient to assume that everyone who performs diagnostic blocks does so in the same way, and would obtain the same results as anyone else. This might be largely true for how needles are introduced and placed, but it is not necessarily so for how responses are obtained and interpreted. Physicians might not agree on what constitutes a positive response.

While so long as there is no published evidence on the reliability of diagnostic blocks for spinal pain, the prospects arise that critics could reject diagnostic blocks for lack of reliability, and payors could deny them for the same reason.

The design of an appropriate study is logistically simple. Obtain the records of 100 consecutive patients who underwent a particular type of block. De-identify the records, and accord each an anonymous identification number. Record the conclusions of the original physician, and redact these from the records. Have second physician read the data in the records, and record a decision as to if the block was positive or not. Calculate the Kappa score for the two physicians.

As a quality control measure, recruit a third and fourth physician. Issue each with a printed checklist of criteria for positive responses, taken from the guidelines for interpreting the block. Have each physician read the 100 records, assisted by the checklist. Calculate the Kappa score for the third and fourth physicians. Calculate the kappa scores for the third and fourth physicians each compared with the readings of the first and second physicians.

The data acquired would indicate if agreement occurs in the realm of conventional practice, if guidelines provide for better agreement, and if practice according to guidelines achieves better agreement than occurs conventional practice.

11. Conclusion

The critical messages this essay are that:

- knowing that a diagnostic test is reliable is necessary for the responsible practice of medicine, and for practice with integrity;
- using tests that lack reliability creates an illusion both to the physician and to their patients;
- Kappa is a classical statistic that can be used to identify diagnostic test of good quality, but more so to identify tests of poor or questionable quality.
- rather than relying on an anonymous equation or calculator, Kappa can be calculated by following simple, logical steps that provide insight into the rationale for the test;
- various mathematical adjustments can be applied to improve the derived values of Kappa, but it is questionable if it is worth doing so for tests that are essentially of moderate quality in clinical practice.
- really good diagnostic tests survive classical analysis for Kappa;
- only questionable tests are brought into relief by unflattering Kappa scores;
- rather than relying on adjustments to Kappa, proponents of questionable tests would be better served by conducting more rigorous studies, or by seeing how to improve the performance of their test or the training in their performance;
- guidelines have been established by which to evaluate and to plan studies of reliability so that their results do not need to be artificially adjusted.

Disclosures

The author was not funded to produce this article, and has no conflicts of interest with its contents.

Appendix 1

The Correlation between Kappa Score and Actual Skill

The value of Kappa has a linear relationship with the actual skills of the observers, but the numerical value of Kappa (if converted to a percentage) is not the same as that of the actual skill. This arises because percentage skill is anchored at zero and 100%, but kappa is anchored at zero for 50% skill, not for zero skill, because 50% skill cannot be distinguished from random guessing.

We can discover how Kappa is related to percentage skill by following some hypothetical examples shown in Appendix Fig. 1. Each table shows what the observed data and Kappa scores would be for observers with increasing degrees of skill. In the first column of Figure Appendix 1 the skills of the observers increase from 50% to 86%. In the second column the skills increase from 90% to 100%.

Skill = 50%		Observer 1		Kappa		Skill = 90%		Observer 1		Kappa
		Pos	Neg					Pos	Neg	
Observer 2	Pos	25	25	0.00		Observer 2	Pos	45	5	0.80
	Neg	25	25				Neg	5	45	
Skill = 60%		Observer 1				Skill = 94%		Observer 1		
		Pos	Neg					Pos	Neg	
Observer 2	Pos	30	20	0.20		Observer 2	Pos	47	3	0.88
	Neg	20	30				Neg	3	47	
Skill = 70%		Observer 1		Kappa		Skill = 96%		Observer 1		Kappa
		Pos	Neg					Pos	Neg	
Observer 2	Pos	35	15	0.40		Observer 2	Pos	48	2	0.92
	Neg	15	35				Neg	2	48	
Skill = 80%		Observer 1		Kappa		Skill = 98%		Observer 1		Kappa
		Pos	Neg					Pos	Neg	
Observer 2	Pos	40	10	0.60		Observer 2	Pos	49	1	0.96
	Neg	10	40				Neg	1	49	
Skill = 86%		Observer 1				Skill = 100%		Observer 1		
		Pos	Neg	Kappa				Pos	Neg	Kappa
Observer 2	Pos	43	7	0.72		Observer 2	Pos	50	0	1.00
	Neg	7	43				Neg	0	50	

Appendix Fig. 1. A set of contingency tables for assessing reliability using Kappa scores, in which the observers have the same skills. The tables show the raw data and Kappa scores that would apply as the Skill of the observers increases from 50% to 100%.

For every table, the prevalence of positive cases is 50%. For a given skill of Z%, Observer 1 correctly identifies Z% of the positive cases and Z% of the negative cases. Observer 2 also correctly identifies those same cases. Thus, the observed agreement (Z%) in each table reflects exactly the mutual skill level of the observers.

Appendix Fig. 1 shows that when the observers have only 50% skill, their Kappa score is 0.00, which is as expected. As their skill increases to 60%, their Kappa score improves, but only to 0.20. As their skill increases to 70%, 80%, and 86%, their Kappa scores improve further, but the numerical values of the Kappa scores remain below the numerical value of their level of skill, although decreasingly so.

Once their skill reaches 90%, their Kappa score (0.80) starts to approximate their skill level. As skill increase beyond 90%, their Kappa scores get progressively closer to their skill level, until their skill level and Kappa score both reach 100%.

These figures show that for very high levels of skill, the value of Kappa understates, but nevertheless reasonably approximates the skill. For moderate levels of skill, the Kappa scores understates the level of skill more greatly but not to an alarming level. Only for low levels of skill is the Kappa score unflattering, if not punitive.

References

[1] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
[2] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
[3] McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012;22:276–82.
[4] Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543–9.
[5] Cicchetti DV, Feinstein AR. High agreement but low kappa, II: resolving the paradoxes. J Clin Epidemiol 1990;43:551–8.
[6] Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46:423–9.
[7] Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther 2005;85:257–68.
[8] Gwet K. Handbook of inter-rater reliability: how to estimate the level of agreement between two or multiple raters. Gaithersburg, MD: STATAXIS Publishing Company; 2001.
[9] Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008;61:29–48.
[10] Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). J Clin Epidemiol 2010;63:854861.
[11] Lucas N, Macaskill P, Irwig L, Moran R, Rickards L, Turner R, Bogduk N. The reliability of a quality appraisal tool for studies of diagnostic reliability (QAREL). BMC Med Res Methodol 2013;13:111.
[12] Engel A, MacVicar J, Bogduk N. A philosophical foundation for diagnostic blocks, with criteria for their validation. Pain Med 2014;15:998–1006.
[13] Engel AJ, Bogduk N. Mathematical validation and credibility of diagnostic blocks for spinal pain. Pain Med 2016;17:1821–8.