The Veterinary Journal 188 (2011) 166–170

Contents lists available at ScienceDirect

The Veterinary Journal

journal homepage: www.elsevier.com/locate/tvjl

Using prevalence indices to aid interpretation and comparison of agreement ratings between two or more observers

Charlotte C. Burn [a,*], Alex A.S. Weir [b]

[a] The Royal Veterinary College, Department of Veterinary Clinical Science, North Mymms, Hertfordshire AL9 7TA, UK
[b] University of Oxford, Department of Zoology, Oxford OX2 7NS, UK

ARTICLE INFO

Article history:
Accepted 14 April 2010

Keywords:
Kappa
Observer agreement
Population homogeneity
Statistics
Subjective scoring

ABSTRACT

Veterinary clinical and epidemiological investigations demand observer reliability. Kappa (κ) statistics are often used to adjust the observed percentage agreement according to that expected by chance. In highly homogenous populations, κ ratings can be poor, despite percentage agreements being high, because the probability of chance agreement is also high. Veterinary researchers are often unsure how to interpret these ambiguous results. It is suggested that prevalence indices (PIs), reflecting the homogeneity of the sample, should be reported alongside percentage agreements and κ values. Here, a published PI calculation is extended, permitting extrapolation to situations involving three or more observers. A process is proposed for classifying results into those that do and do not attain clinically useful ratings, and those tested on excessively homogenous populations and which are therefore inconclusive. Pre-selection of balanced populations, or adjustment of scoring thresholds, can help reduce population homogeneity. Reporting PIs in observer reliability studies in veterinary science and other disciplines enables reliability to be interpreted usefully and allows results to be compared between studies.

Introduction

Subjective classification systems are often used to assess animal health or behaviour when objective measures are either lacking or are too invasive, intrusive or logistically unmanageable for the particular situation being investigated. When using subjective scoring systems, assessing the extent to which different observers agree on how subjects (which in veterinary science might be animals or clinical samples) are classified represents an important part of the validation process. Moreover, large-scale epidemiological and longitudinal studies often rely on data being collected by more than one observer (Waters et al., 2002; Dawkins et al., 2004; Rutherford et al., 2009; Burn et al., 2010). Adequate consistency between these observers is critically important if the data are to be comparable and representative of the populations sampled.

Kappa (κ) statistics are used to assess the extent to which the proportion of agreement within or between observers is better than by chance. In this way, κ statistics are more stringent than correlations or raw percentage agreements (% agreements) alone (Hoehler, 2000). Also, κ statistics can be used to assess consistency and agreement between alternative testing methods for measuring the same variable, such as equivalent methods for diagnosing a disease. The word 'observer' here can therefore refer to any observational means, whether a human observer, a measuring instrument or a technique.

Finding good agreement, as indicated by % agreements close to 100 and by κ values close to 1, allows measurements to be made by different observers (or by the same observer at different time points) with some confidence in their consistency. Equally, finding poor observer agreement (% agreements and κ values both close to 0) is useful for alerting users to scoring systems or diagnostic methods that require modification, clearer definition or more in-depth training. However, when the % agreement is high but the κ value is low, researchers can be unsure as to how to proceed.

In this paper, we show that this ambiguity can be understood in the context of prevalence imbalance in the sample population and we offer approaches for overcoming this issue and for classifying results in a practical manner. We only consider reliability for binary scoring systems, because analysing multiple categories simultaneously risks attaining good overall agreement despite a minority category being frequently misclassified. Kraemer and colleagues (2004) advise that nominal variables with more than two categories should be broken down into their binary components, so that each category can be independently tested against the other categories combined.

* Corresponding author. Present address: Department of Clinical Veterinary Science, The Royal Veterinary College, Hatfield AL9 7TA, UK. Tel.: +44 1707 666000; fax: +44 1707 652090.
E-mail address: cburn@rvc.ac.uk (C.C. Burn).

1090-0233/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tvjl.2010.04.021

Relationship between population prevalence imbalances and κ values

The equations for κ are described and explained in detail by Fleiss (1971), but the overall principle is:

$$\kappa = \frac{P(observed) - P(chance)}{1 - P(chance)} \quad (1)$$

where $P(observed)$ is the proportion of subjects that the observers agree on, $P(chance)$ is the proportion of agreement expected by chance and 1 is the maximum possible agreement. $P(chance)$ depends on the proportion of assignments to the different categories (both agreements and disagreements). For Fleiss' κ this is calculated as the sum of the squared proportions of assignments to each category (Fleiss, 1971). For example, with 10 subjects, two categories and two observers, if the two observers together made 14 assignments to category a (and therefore 6 to category b), $P(chance)$ would be $(14/20)^2 + (6/20)^2 = 0.58$ (Table 1).

Some authors have proposed a 'prevalence index' (PI) to describe the balance of the two categories in the population being rated. For two observers using a binary scoring system, PI is calculated as follows (Byrt et al., 1993; Sim and Wright, 2005):

$$\text{PI} = \frac{|a - d|}{n} \quad (2)$$

where a is the number of agreed upon subjects in one of the categories, d is the number of agreed upon subjects for the other category and n is the total number of possible agreements (i.e. the number of subjects). Thus, it is the absolute difference between the numbers of subjects that both observers agreed on in each category divided by the total number of subjects. For a PI that can be extrapolated to more than two observers, we propose a new calculation described below, but the interpretations remain the same as for the two-observer calculation. With either calculation, when the PI is 0, this indicates no imbalance (50% of agreements fall into one category and 50% into the other), while a PI of 1 indicates that all agreements fall into one category.

Table 1
Hypothetical examples with the same PI and $P(chance)$, but differing agreement and κ.

Subject ID	Number of allocations to category a	Number of allocations to category b
(a)		
1	2	0
2	2	0
3	2	0
4	2	0
5	2	0
6	2	0
7	2	0
8	0	2
9	0	2
10	0	2
Total allocations	14	6
Subject ID	Number of allocations to category a	Number of allocations to category b
(b)		
1	2	0
2	2	0
3	2	0
4	2	0
5	1	1
6	1	1
7	1	1
8	1	1
9	1	1
10	1	1
Total allocations	14	6

In both situations, two observers together made 14 assignments to category a (and therefore six to category b); for the Fleiss κ calculation $P(chance)$ is $(14/20)^2 + (6/20)^2 = 0.58$. The PI is 0.4, regardless of whether it is calculated using Eq. (2) [for (a): $(7 - 3)/10 = 0.4$; and for (b): $(4 - 0)/10 = 0.4$] or Eq. (3) [for both tables: $(14 - 6)/20 = 0.4$]. However, in (a) there is 100% agreement and $\kappa = 1$; while in (b) there is 40% agreement and $\kappa = 0$.

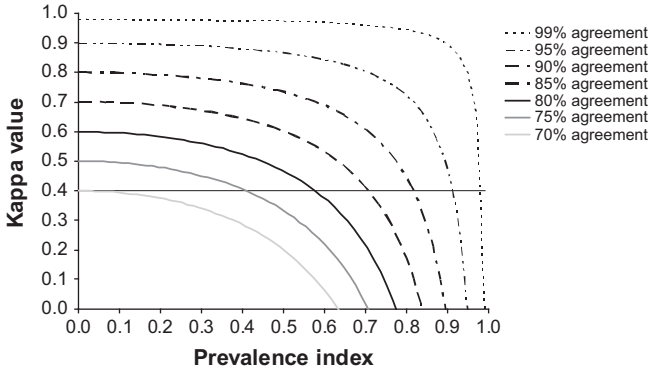

Fig. 1. Relationships between prevalence indices and Fleiss kappa values for a range of percentage agreements. The solid line at $\kappa = 0.4$ shows a commonly used minimum threshold for clinical relevance (Sim and Wright, 2005).

For any given % agreement, as the PI increases, κ decreases. This is because the probability of agreeing purely by chance is very high in near-homogenous populations, as when two observers both know that a disease is rare (Elbers et al., 2004) or, if they frequently fail to detect the clinical signs, they may agree that most animals are healthy purely by chance and yet their agreement would offer no assurance that they could consistently identify animals with the disease. Thus, in near-homogenous populations, evidence for agreement above chance levels is difficult or impossible to identify and this is reflected in a low κ value (Hoehler, 2000; Vach, 2005). For given % agreements, Fig. 1 illustrates the relationship between Fleiss κ values and population homogeneity as measured using PIs (regardless of the number of observers).

The dependence of κ on the prevalence of the categories being assessed has sometimes been treated as an undesirable limitation or a 'paradox' of κ statistics (Byrt et al., 1993; Kundel and Polansky, 2003; Randolph, 2005). This has led some authors to prefer the raw positive and negative agreements alone (e.g. for questionnaire data: Sargeant and Martin, 1998) or to implicitly accept a lower κ when the PI is known to be high (e.g. for assessing behavioural responses of cattle to humans: Rousing and Waiblinger, 2004).

Other authors have proposed alternative κ calculations, such as 'prevalence adjusted bias adjusted κ' (PABAK) (Byrt et al., 1993) or 'free-marginal multirater κ' (Randolph, 2005), but these approaches have been criticised for readjusting for the very factors that κ is designed to control for (Hoehler, 2000). In fact, κ tests are a useful tool for assessing whether high % agreements are likely to be due to genuinely consistent observations, or whether they could be a consequence of an unbalanced sample population.

As a side note, some statistical packages provide P values to indicate whether agreement is above 'chance', but these significance tests are usually based on a null hypothesis of observed agreement being 50%, which is when $\kappa \leqslant 0$ (Sim and Wright, 2005). Therefore they are of limited use, being non-significant only when agreement is extremely poor; κ is more informative.

Calculating prevalence indices for multiple observers

The discussion to date about issues of prevalence and κ statistics has largely focussed on situations involving two observers. When multiple observers are compared, Fleiss' κ calculations are usually used instead of Cohen's κ, because they can weight the