Epidémiol. santé anim., 1997, 31-32

HOW GOOD ARE DATA FROM QUESTIONNAIRES : A REVIEW OF MEASURING RELIABILITY

Slater M.[1]

Depuis que les données sont très souvent récoltées à partir d'un questionnaire en vue de déterminer l'exposition et le statut vis-à-vis d'une maladie, elles sont sujettes à des critiques. La fiabilité des données d'un questionnaire doit être évaluée avant toute estimation de leur validité. La fiabilité est définie généralement par la possibilité d'avoir la même information sur le même individu à différents moments ou par différents investigateurs. Il existe deux types de fiabilité : la cohérence interne et la stabilité. Cette présentation se limitera à la stabilité, et plus précisément la méthode du « test-retest ». La stabilité reflète la capacité du questionnaire à collecter les données en différentes occasions.

Des mesures spécifiques pour estimer la stabilité sont classées en deux groupes : celles relatives aux variables continues (coefficient de corrélation), et celles pour les variables catégorielles (kappa). Pour les données continues, suivant une distribution normale, le coefficient de corrélation de Pearson (PCC) est le plus fréquemment utilisé. Le coefficient de corrélation de Spearman (SCC) est plutôt utilisé pour les distributions non normales, ceci est équivalent à l'utilisation du PCC sur les rangs des données continues. Le coefficient de corrélation intra-classe (ICC) est recommandé comme plus approprié que le PCC, pour tester la fiabilité lorsque les données sont continues. Concernant les données ordinales avec peu de catégories, le test Kappa est conseillé. Le test Kappa pondéré (quadratique) est équivalent au ICC. Pour les données nominales, le Kappa est la seule option.

INTRODUCTION

We use questionnaires extensively in epidemiology to collect data. One very important issue is the quality of the data collected. This is generally assessed by means of reliability and validity. This paper will address reliability. Assessment of reliability should be done before validity since data which cannot be obtained in a consistent manner cannot possibly be accurate. Reliability may be thought of as way of quantifying the amount of random and systematic error in the questionnaire (Streiner and Norman, 1995). Another way of defining reliability is the ratio of the variability (eg, variance) between subjects to the total variability (from subjects, observers, random error, etc.) of the measurement. In the context of questionnaire evaluation, reliability is generally defined as the ability to obtain the same information from the same individuals at different times (test-retest) or by different observers/interviewers (inter-observer), for example. I will focus on test-retest methods in this paper but many principles are similar for other types of assessment methods.

TYPES OF RELIABILITY

In general, there are two types of reliability that can be assessed: internal consistency and stability (Streiner and Norman, 1995). Internal consistency is measured using a single administration of the instrument when there are multiple questions addressing the same underlying issue or problem. Commonly used tests are Cronbach's alpha and split-halves. These are frequently found in psychology. In many veterinary epidemiologic situations, I have not found that we have multiple questions addressing the same issue, making this a difficult type of reliability to assess.

Stability reflects the ability of the questionnaire to collect data on more than one occasion. Some common permutations are inter-observer reliability (two different observers at the same time), intra-observer (the same observer at different times) and test-retest (the same subjects evaluated some time period apart). It is sometimes confusing to sort out observers, interviewers and subjects. For example, in a telephone interview done one month apart by the same interviewer of the same dog owner, the method of reliability evaluation is generally accepted to be test-retest and yet there could also be error due to intra-observer (the interviewer) variability.

SOURCES OF VARIABILITY

The administration method of the questionnaire will partly determine the potential source(s) of variability. For self-administered questionnaires, the subjects are the "observers" and only variability due to random or systematic error in the questions would be assessed. For questionnaires with interviewers, the interviewers may be considered to be "observers", adding another potential source. However, depending on the situation, the interviewer error may not be important.

Another factor in determining the source of the variability is whether you are assessing the reliability of a single item (eg, a 1-10 pain scale), a summary scale from a set of items (eg, an overall lameness assessment from multiple questions), or an average of repeated assessments (eg, average daily protein intake from two interviews at different times). I will discuss some of these implications later in the paper.

[1] Department of Veterinary Anatomy and Public Health, College of Veterinary Medicine, Texas A&M University, College Station, TX 77843-4458, USA

TEST-RETEST RELIABILITY

This method is what is usually found in the companion animal epidemiologic literature (Sonnenschein 1988; Slater et al., 1992; Slater et al., 1995; Reeves et al., 1996). Typically, the same questionnaire is administered to the same subjects (usually the owners of the animals being studied) some time interval apart. The subjects are asked to provide information about the same time period in both interviews. In some cases, the same interviewers are used for both interviews (eliminating different "observers" as a source of error), but that is not always possible (hence the importance of interviewer training and monitoring to decrease interviewer bias).

The most difficult aspect of the test-retest method is determining the appropriate time interval between interviews (and then actually getting the interviews done on time). One consideration is the time period of interest to the investigator. In human nutritional epidemiology, if the past year's intake is of interest, then intervals of up to several months should be fine and differences between interviews should be due to the questionnaire and not to recent changes in diet (Willett, 1990). For a general recommendation, Streiner and Norman (1995) suggests two to 14 days. Another suggestion is at least one month apart (Armstrong et al., 1992). In the companion animal literature cited above, one month is commonly considered long enough for owners not to remember their answers but short enough for events to have remained the same. The exact interval will depend on: 1) how complicated the information is (if more complex, people won't remember it very easily and shorter intervals will work); 2) the time period of interest for the investigator (longer periods are likely to be more stable and could use longer intervals); 3) length of the questionnaire (longer questionnaires, like complex information, make it hard for owners to remember previous answers); and 4) type of information (since data that change very slowly should be able to be accurately assessed across very long intervals).

MEASURES OF TEST-RETEST RELIABILITY

These measures can be roughly divided into two types: those for continuous variables (correlation coefficients) and those for categorical variables (kappas). Guidelines for interpretation are somewhat controversial and will depend on the circumstances. However, general suggestions are: < 0.4, poor agreement; 0.4 - 0.75, fair to good agreement; > 0.75 excellent agreement (Fleiss, 1981; Willett, 1990). For continuous, normally distributed data, Pearson correlation coefficient (PCC) is very commonly used. Recall that PCC measures the strength of the linear association between the two continuous variables. Therefore, PCC will equal one (a perfect correlation) when the points form a perfect line (Streiner and Norman, 1995). Spearman correlation coefficient (SCC) is interpreted similarly and is suitable for skewed or non-normal data (Willett, 1990) and is very similar to doing a PCC on the ranks of the continuous variable. It can also be used for ordinal data (I generally use it for four or five or more categories) (Armstrong et al., 1992).

The intraclass correlation coefficient has been recommended as a more appropriate measure of reliability for continuous variables than the PCC for several reasons (Streiner and Norman, 1995). Firstly, it measures exact agreement and will only be equal to one if each subject has identical observations. Secondly, if you are comparing more than a pair of interviews (eg, three or more), a single ICC can be used rather than multiple pairwise PCCs. Having said all this, in practice the ICC and the PCC are usually similar (since the ICC is more conservative it will generally be slightly lower than the PCC). Further, the PCC is commonly computed directly by statistical packages and the ICC must be calculated from the appropriate analysis of variance table. For example, in a subset of a test-retest study of dog nutrition, the ICC for usual daily protein intake = 0.78, the PCC = 0.78 and the SCC = 0.73, all indicating very good agreement. These data were somewhat skewed.

Another thing to consider when using the ICC which is both a strength and, in my opinion, a weakness is that there are multiple versions of it, nicely presented by Streiner and Norman (1995). The simplest version is the ratio of the variance between subjects to the variance between subject plus the error variance. This is the ICC I reported for the previous example and is probably appropriate for test-retest data (Armstrong et al., 1992). It does not, however, take into account the variance from the two interviews (which could be considered as two "observers"). In this case, this variance would be added into the denominator. Fortunately, this is a small quantity and the new ICC = 0.77, a negligible difference in this situation. This confusion is also less problematic, if the same interviewer does both interviews. To further complicate the issue, we have also not taken into account the variance due to different interviewers if such is the case. This is also considered a source of variability due to "observer". The suggested guidelines are that "observer" does not need to be included (treated as a fixed effect) if the same observers in the reliability study will be used in the main study. If, however, the "observers" are considered a random sample of possible observers (random effect) then their variance should be included in the denominator. I have not been able to find a reference that specifically addresses the problem of test-retest with interviewers.

Additional situations in which "observer" is considered a fixed factor are: when a single item scale is evaluated (all subjects complete it) and, similarly, when multiple items become a single summary measurement (all subjects complete all items) (Armstrong et al., 1992; Streiner and Norman, 1995). The latter does occur in the test-retest situation.

For ordinal data for two to four (or so) categories, Kappa is suggested (Armstrong et al., 1992; Streiner and Norman, 1995). Kappa measures the amount of agreement beyond what you would expect due to chance and is therefore preferred over percent agreement. For two categories, a single unweighted kappa is clearly appropriate. For three categories, an overall unweighted kappa is useful; it is usually helpful to compare each category in a pairwise fashion to see which ones are the source of disagreement since the first and third categories of an ordered variable are usually more easily discriminated among than two adjacent categories (Maclure and Willett, 1987). A weighted kappa may also be used for variables with three or more categories. A weighted kappa gives partial credit for agreement that is close but not exact (Armstrong et al., 1992). Unfortunately, the issue then becomes: what weights do you use? A helpful comparison of linear vs quadratic weighting is presented in Brenner and Kliebsch (1996).

Since using quadratic weighting results in the kappa statistic being equivalent to the ICC, this fairly standard weighting system is recommended (Streiner and Norman, 1995).
Fortunately, for nominal data, kappa, weighted or unweighted is the only option. There are three other general points to consider when using kappa. Kappa is affected by the number of categories since it measures exact agreement. Fewer categories will give a higher kappa. On the other hand, weighted kappas (particularly using quadratic weights) gives a lower kappa with fewer categories (Maclure and Willett, 1987; Brenner and Kliebsch, 1996). Kappa is also dependent on the distribution of the exposure in the population you are studying (Armstrong et al., 1992). Kappa should not be used on a continuous variable that has been arbitrarily categorized (Maclure and Willett, 1987).

CONCLUSIONS

Evaluating the reliability of your questionnaire is critical if you are to trust the results of the study. Test-retest methods are most commonly used in companion animal epidemiology and frequently involve telephone interviews. One of the correlation coefficients is most suitable to assess reliability for continuous variables; kappa should be used for nominal variables and for ordinal variables with few categories. The ICC is considered the most appropriate measure of reliability; it has many variations which allow you to explore different sources of variability but must be calculated from the analysis of variance table. In many circumstances, PCC will be very similar, differing only by a small amount of random error. Weighted kappa, while sensitive to the number of categories, is equivalent to the ICC if quadratic weights are used.

BIBLIOGRAPHY

Armstrong B.K., White E., Saracci R., 1992. Principles of exposure measurement in epidemiology, Monographs in epidemiology and biostatistics volume 21. Chapter 4. Oxford University Press, New York.

Brenner H., Kliebsch U., 1996. Dependence of weighted kappa coefficients on the number of categories. Epidemiology 7, 199-202.

Fleiss J.L., 1981. Statistical methods for rates and proportions. John Wiley & Sons, New York.

Maclure M., Willett W.C., 1987. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 126, 161-169.

Reeves M.J., Salman M.D., Smith G., 1996. Repeatability of equine health and management information obtained via a telephone questionnaire. Prev Vet Med 26, 347-351.

Slater M.R., Scarlett J.M., Donoghue S. Erb H.N., 1992. The repeatability and validity of a telephone questionnaire on diet and exercise in dogs. Prev Vet Med 13, 77-91.

Slater M.R., Robinson L.E, Zoran D.L., Wallace K.A., Scarlett JM., 1995. Diet and exercise patterns in pet dogs. J Am Vet Med Assoc 207(2), 186-189.

Sonnenschein E.G., 1988. A case-control study of nutritional factors and spontaneous breast cancer in pet dogs. UMI Dissertation Information Service, Ann Arbor, MI. Order number 8816237.

Streiner D.L., Norman G.R., 1995. Health measurement scales: a practical guide to their development and use. Oxford University Press, New York.

Willett W.C., 1990. Nutritional Epidemiology, Monographs in epidemiology and biostatistics volume 15. Oxford University Press, New York.