Commentary

No significant difference: use of statistical methods for testing equivalence in clinical veterinary literature

Robert M. Christley, BVSc, MVetClinStud, PhD, and Stuart W. J. Reid, BMVS, PhD

Clinical veterinary research often uses statistical methods to investigate whether one factor (eg, an agent, method, or treatment) alters an outcome. Therefore, it is common for researchers to report that one treatment resulted in a significantly greater number of animals surviving than another treatment or no treatment. If the study is correctly designed and the statistical analyses are conducted appropriately, this is a perfectly legitimate claim. However, problems may arise when researchers report results that are not significantly different. By convention, the significance level is set at 5%, and a P value > 0.05 is termed not significant.[1] In these circumstances, it is tempting to report that one factor is comparable to, equivalent with, or not different from the second factor. The implication, either implicit or explicit, is that the factors being considered are the same or have the same effects. Actually, all that has been demonstrated is an absence of evidence of a difference.[1] These statements are quite different, and concluding that two factors that are not significantly different are comparable could lead to erroneous conclusions in relation to the research question. If we believe that two treatments are not going to result in exactly the same response, the correct question is whether or not the difference between the treatments is sufficiently small enough that they could be used interchangeably.

Commonly, nonsignificant differences are interpreted as indicating equivalence. A review[2] of medical literature revealed that the conclusions of 67% of studies claiming clinical or therapeutic equivalency were based on a failure to detect significant differences. Only 23% of studies confirmed equivalence with appropriate statistical testing, and 10% performed no statistical analysis. There are numerous examples in the veterinary literature of equivalence of two factors being based on lack of a significant difference. When conclusions are based on the inability to reject the null hypothesis of no significant difference, such conclusions may be erroneous, particularly when the sample size is small. Such results are also not very informative. Therefore, when the purpose of a study is to determine that two sets of results are equivalent, useful, yet similar, questions are: is there evidence in the data that the results are equivalent, and, given that the two groups of results are unlikely to be identical, is the difference sufficiently small to be considered negligible?

Statistical methods[3-8] to provide evidence for equivalence between different treatments, methods, or other factors have been described. Primarily, these have been developed for the pharmaceutical industry to demonstrate that a new formulation of a drug is equivalent to an existing formulation. Initially, concern was limited to detecting average or typical equivalence; that is, determining whether the typical effect of two or more drugs is similar across a population. However, recently this concept has been disaggregated to equivalence at the level of the population and the individual.[9,10]

Several approaches may be used to investigate equivalence or test for the degree of similarity. These include hypothesis testing, confidence intervals, and graphic means. However, such methods are rarely used in veterinary clinical research and may seem daunting to researchers without a background in statistics. The purpose of this commentary is to illustrate some of these methods with examples from the veterinary clinical literature. It is important to note that we are not suggesting in any instance that the original authors intended to mislead. Rather, our intention is to show that the use of alternative statistical methods might allow a more detailed interpretation of the results; indeed, we are grateful to the authors of the articles we have selected for publishing their studies in sufficient detail to allow further analysis.

Selection of Research Articles

Articles were selected by review of recent issues of several journals that publish veterinary clinical research. The search procedures were unstructured and not exhaustive. The objective was to identify research articles in which important conclusions had been made on the basis of nonsignificant differences detected by use of standard statistical procedures. In particular, articles in which sufficient data were presented to enable reanalysis, using the methods of interest, were sought.

From Comparative Epidemiology and Informatics, Institute for Comparative Medicine, Faculty of Veterinary Medicine, University of Glasgow, Bearsden, G61 1QH, UK (Christley, Reid); and the Department of Statistics and Modelling Science, Faculty of Science, University of Strathclyde, Glasgow, G1 1XH, UK (Reid). Dr. Christley's present address is the Epidemiology Group, Department of Veterinary Clinical Science, Faculty of Veterinary Science, University of Liverpool, Leahurst, CH64 7TE, UK.
Address correspondence to Dr. Christley.

Statistical Methods

Published methods of performing equivalence testing were used. These included plotting the mean versus the difference of two measures,[8,11] the two one-sided tests for normally distributed data,[5] and the Hauck-Anderson corrected two one-sided tests for proportions.[12] These were performed with a spreadsheet program.[a] The programming was cross-checked for errors via manual calculation and by comparison with published results. Some of the spreadsheets are available at www.vie.gla.ac.uk/equivalence/. The formulae used were listed (**Appendices 1 and 2**).

Determining Equivalence of Two Tests or Observers

Studies are often designed to investigate whether results obtained from two instruments or observers agree sufficiently to be used interchangeably. Methods that have been used include paired *t* tests,[13] correlation coefficients,[14] and plotting the difference between the methods against the standard method.[15] Although each of these methods may have a role in the interpretation of the results, incorrect interpretation may be misleading.[11,16] For example, evidence of a clinically important statistical difference detected by use of a paired *t*-test or a small correlation coefficient provides evidence against equivalence. However, a large correlation coefficient or the absence of a significant difference does not provide sufficient evidence to claim equivalence. An appropriate method to provide evidence for equivalence for each pair of data has been suggested by Bland and Altman[8,11,16] and involves plotting the difference between results obtained by these methods against the mean of the results of the methods.

As an example, consider a comparison of lameness ratings in sheep by two observers.[13] For each of 45 sheep, the two observers rated the degree of lameness with a **visual analogue scale (VAS)**, which scored the lameness from 0 (no lameness) to 100 (could not be more lame). In addition to using descriptive statistics, the authors concluded that lameness score obtained with the VAS was reproducible because of failure to identify a significant difference between the observers with a paired *t* test (our calculation, $P = 0.8$), essentially comparing the mean of the differences against zero.

Another approach would be to plot the results of one observer against the other (**Fig 1**) and calculate the correlation coefficient ($r = 0.92$; $P < 0.001$). However, this high correlation does not mean the two methods agree.[16] The correlation coefficient measures the strength of the relationship between the two sets of observations, not the agreement between them. In other words, there will be a perfect correlation if the points lie along any straight line, even if they were measured on different scales. For example, halving the results of observer 2 does not alter the correlation coefficient.

Bland and Altman[11] recommend an alternative method for the analysis of such data, which is based on simple graphic techniques and elementary calculations. Their approach involves plotting the difference between each pair of observations against the mean of that pair of observations. If the measurements made by each observer are exactly equivalent, the data points should lie along the line of zero difference, regardless of the mean measurement. When applied to the lameness data (**Fig 2**), there is evidence of considerable lack of agreement between the two observers with discrepancies of up to 25 units. Although inspection of Figure 1 suggests there are discrepancies between the methods, Figure 2 clearly illustrates the magnitude of these differences. The lack of agreement can further be summarized by calculating the 95% limits of agreement. For the VAS data, the mean difference between observers is –0.4, and SD is 11.4. We would expect that 95% of observations would lie between 0.4 – 1.96 SD and 0.4 + 1.96 SD (ie, the limits of agreement), assuming that the differences follow a normal distribution. In fact, the differences are likely to follow a normal distribution, because we have removed much of the variation between subjects and are left with the measurement error. The measurements themselves do not need to follow a normal distribution, and often they will not.

Provided that the differences within the limits of agreement were not clinically important, the observations by the two observers could be used interchangeably. For the VAS data, the limits of agreement are –22.7 to 21.9. Therefore, one observer may assign scores that are 23 units less than or 22 units greater than those assigned by the other observer. Whether such a range of difference (45 units) is important is a clinical, rather than statistical, question. However, the limits of agreement suggest a lack of agreement that is not immediately obvious from the correlation coefficients or Figure 1.

This method is more informative than reporting no significant difference on the basis of a paired *t* test.

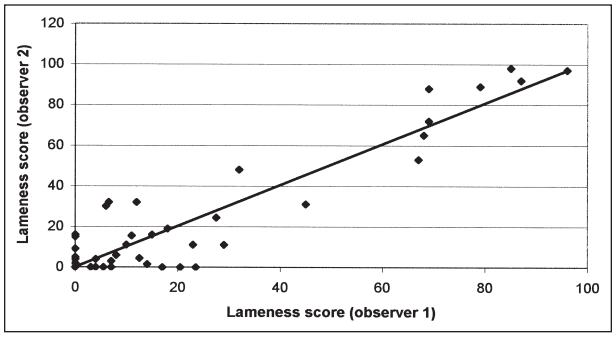

Figure 1—Correlation between two observers' measurements of lameness scores with a visual analogue scale in 45 sheep.[13] Solid line indicates the line of equality.

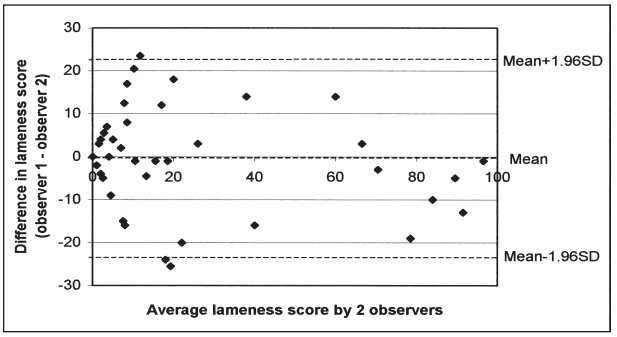

Figure 2—Difference versus mean of each pair of values for lameness in sheep measured by two observers with a visual analogue scale, with 95% limits of agreement[13] (upper and lower dotted lines).

The limits of agreement indicate the spread of the data. If the authors or readers believe that these limits are not excessively wide, equivalence can be the conclusion. The plot of the difference against the mean also provides information regarding the nature of the relationship between the observations. In this example, the relationship appears reasonably uniform. However, in other instances, we may observe a nonuniform relationship, with the difference increasing or decreasing as the mean increases. In addition, if the difference between the observers was considerably different from zero, there is likely to be measurement bias between the observers. If this bias is consistent, it is a simple matter to adjust for it by subtracting the mean difference from the measurements made by the second observer. This is not a problem in this example.

Extensions of this method are available for data in which the relationship between the difference and the mean difference is not uniform, in which there are repeated measurements, and in which the differences do not have a normal distribution.[8]

Detecting Equivalence of Outcomes after Two Interventions

Proportions—Many studies investigate the effect of an intervention (such as drug administration or surgery) by measuring a clinical outcome. Often, the results are in the form of proportions; for example, the proportion of animals that recovered or survived. It may be of interest to determine whether the intervention does not alter the likely outcome, compared with a reference group. Several comparisons may be made in order to analyze such data; the difference in success probabilities is usually considered, although the ratio of the proportions (relative risk) and the odds of the rates of the outcome have also been used.[3,6,12,17,18] Because of problems with the specification of the acceptance interval, the use of relative risk is not recommended.[18] Similarly, the use of the odds ratio is not widely reported and may be undesirably conservative.[18] Therefore, when appropriate, using the difference in probabilities is recommended.

Several methods for assessing equivalence on the basis of the difference in probabilities are available.[3,6,12,17,19-21] These differ in relation to the calculation of the SD, correction factor, or both. Despite resulting in some loss in statistical power, the Hauck-Anderson corrected classical procedure has been recommended[12,17] (Appendix 1).

This procedure can be used in two ways. First, it can be used to test the hypotheses of equivalence between two proportions, and, second, it can be used to create confidence intervals of the difference between the proportions. We believe that the confidence interval approach is more intuitive and informative.

In a study of the effects of early gonadectomy (prepubertal [prior to 24 weeks of age]) in cats, one of the outcomes of interests was the rate of behavioral problems.[22] Of 75 cats that underwent traditional-age gonadectomy, 26 (35%) developed behavioral abnormalities. In contrast, 49 of 188 (26%) cats that underwent prepubertal gonadectomy developed behavioral problems. The authors correctly concluded that there was no significant difference between the rates of behavioral problems in these groups (our calculations, $P = 0.2$). The authors suggest that "prepubertal gonadectomy may be performed safely in cats without concern for increased incidence of …behavioral problems." However, the finding of lack of significant difference does not mean the proportion of animals with behavioral problems is the same in each group.

From the information presented in the article, further hypotheses about the effect of early gonadectomy can be investigated. For example, we could consider whether the difference between the probabilities of behavioral problems is greater than or less than specified lower and upper limits, respectively. That is,

$$H_0 : P_T - P_R \le \theta_1 \text{ or } P_T - P_R \ge \theta_2 \text{ versus } H_1 : \theta_1 < P_T - P_R < \theta_2$$

where P_T and P_R are proportion of behavioral problems in the treatment (ie, early gonadectomy) group and reference group, respectively, θ_1 and θ_2 are the prespecified lower and upper limits for equivalence, and $\theta_1 < 0 < \theta_2$. This can be rewritten as two one-sided hypotheses

$$H_{01} : P_T - P_R \le \theta_1 \text{ versus } H_{11} : P_T - P_R > \theta_1$$
$$H_{02} : P_T - P_R \ge \theta_2 \text{ versus } H_{12} : P_T - P_R < \theta_2$$

We could then investigate whether the study provides evidence of equivalent rates of behavioral problems, given the prespecified limits. If we reject both of the null hypotheses, we can conclude that the proportions are equivalent, within the acceptance limits θ_1 and θ_2. In this example, if the only concern was to determine that early gonadectomy did not increase behavioral abnormalities, we could just consider H_{02}, that early gonadectomy does not increase the rate of behavioral abnormalities by more than θ_2. However, it may also be of interest to investigate whether early gonadectomy decreases the rate of behavioral problems by θ_1, so H_{01} is also considered here. Note also that asymmetric limits can be used (eg, –0.1 and 0.3), but for simplicity, symmetric limits will be used in the following example.

Let us assume that a group of experts conclude that, because early gonadectomy is an important means to control the companion animal overpopulation problem, a difference in the rate of behavioral problems of 0.2 (20%) would be acceptable. Using the Hauck-Anderson corrected classical method (Appendix 1),[12] the P value for H_{01} is 0.047 and for H_{02} is < 0.001. Therefore, we can conclude that, within the acceptance limits, the proportions are equivalent.

However, expert opinion often differs. For example, other experts reading the article might believe that, because behavioral problems are a major cause of animals being surrendered to animal shelters,[23] increasing the rate of behavioral problems by 20% may be unacceptable. Hence, these experts may be unsatisfied with the conclusions of the article. Determination of the confidence interval for the difference between the proportions removes the specification of the equivalence limits and enables estimation of the interval within which the true difference is likely to be. In this way, readers can determine whether or not they consider

the maximum differences, in either direction, to be sufficiently small to consider the proportions equivalent. In this example, the 90% confidence interval is –0.2 to 0.03. Hence, rather than merely concluding that the result was not significant, we could conclude that early gonadectomy does not increase the incidence of behavioral problems by more that 3% and may actually decrease it by as much as 20% (90% confidence interval). In this way, there is no need to consider whether the proportions are equivalent, just whether the range of possible differences would be acceptable. We do recommend, however, that such considerations be made prior to analysis.

Continuous data—Frequently, studies are designed to determine the equivalence of interventions, in terms of a continuous outcome (such as a physiologic or biochemical parameter or serum concentrations of a drug). As with proportional data, several methods are available to investigate such equivalence, including hypothesis testing and the use of confidence intervals (limits of agreement). The hypothesis testing approach involves construction of a null hypothesis that the difference between the means is within a certain small range[5] (Appendix 2). That is,

$$H_0 : \mu_T - \mu_R \leq \delta_1 \text{ or } \mu_T - \mu_R \geq \delta_2 \text{ versus } H_1 : \delta_1 < \mu_T - \mu_R < \delta_2$$

where δ_1 and δ_2 are the prespecified lower and upper limits, respectively. Once again, this can be rewritten as two one-sided hypotheses

$$H_{01} : \mu_T - \mu_R \leq \delta_1 \text{ versus } H_{11} : \mu_T - \mu_R > \delta_1$$
$$\text{and}$$
$$H_{02} : \mu_T - \mu_R \geq \delta_2 \text{ versus } H_1 : \mu_T - \mu_R < \delta_2$$

The procedure for testing these hypotheses involves two one-sided tests (one for each component of the null hypothesis).[5] Detection of a sufficiently small *P* value (usually $P < 0.05$) for each of these tests would allow rejection of the null hypothesis and the conclusion that the difference between the means is between δ_1 and δ_2.

As an example, Ko et al[24] investigated the effect of carprofen on the **glomerular filtration rates (GFR)** of healthy dogs after anesthesia and concluded that there was no significant difference from baseline values. This conclusion is statistically correct. However, because only five dogs were assigned to each of the treatment and reference groups, there may be concern about the statistical power of the study to detect a difference, if present. The study design allowed several comparisons, including treatment versus control after the first anesthesia (ie, no intervention to either group), treatment versus control after the second anesthesia (ie, effect of carprofen compared with control), treatment group compared after first and second anesthesia (paired comparison of treatment effect), and control group after first and second anesthesia (paired comparison of control effect). The authors correctly used ANOVA with posthoc tests to make these comparisons. Multiple testing procedures to declare equivalence are reported in the literature but are outside the scope of this commentary.

In the study by Ko et al,[24] mean ± SD GFR values were 2.16 ± 0.45 and 1.98 ± 0.83 mL/kg/h for the treatment and control groups, respectively. Therefore, the pooled SD is 0.67.[25] With this information, we can investigate whether the difference between the means of the control and treatment groups is within certain limits. We could consider that carprofen should not alter the GFR by > 30% of the control value. That is, δ_1 and δ_2 equal –0.6 mL/kg/h and 0.6 mL/kg/h, respectively. Testing these hypotheses indicates that carprofen is unlikely to reduce GFR by 30% ($P = 0.03$), but we cannot conclude that it does not increase GFR by 30% ($P = 0.2$).

Alternatively, we could calculate the maximum values of δ_1 and δ_2 that would occur with $P = 0.05$. In the carprofen example, the lower and upper limits of the difference between the means are –0.51 mL/kg/h and 1.24 mL/kg/h, respectively. Therefore, although it appears that, if anything, carprofen increased GFR (by as much as 1.24 mL/kg/h; ie, 63% greater than the control mean), the results of the study indicate that carprofen could reduce GFR by as much as 0.51 mL/kg/h (a 26% decrease from control values). The question now is whether decreasing GFR by one-quarter or increasing it by two-thirds is of clinical concern. If either of these is potentially important, then it cannot be concluded that carprofen does not affect GFR to an important degree, and further study, possibly with increased sample size, is warranted.

Conclusions

We have illustrated some of the analytic methods available for providing evidence that two factors are equivalent. These methods include graphic and quantitative approaches. The methods have advantages over standard statistical procedures, because they allow a conclusion of equivalence (within specified limits) rather than simply a lack of significant difference and are, therefore, less open to misinterpretation. In addition, they can be used to determine a confidence interval for the difference between groups. Such approaches should be considered when the hypothesis is of no difference or when no significant difference is detected.

One useful aspect of the hypothesis testing and confidence interval approaches for normally and binomially distributed data is that they can be applied retrospectively, using details typically provided in published articles. For normally distributed continuous data, information regarding mean, variance (or SD), and number of animals is required. For binomial data, the number of animals with and without the outcome in the treatment and reference groups is required. However, these methods are only appropriate for certain study designs; many other analytic techniques are available for other types of studies. The methods reported here are simple, whereas other methods may be considerably more complex. Advice from a statistician is recommended before such analyses are undertaken.

[a]Microsoft Excel 2000, Microsoft Corp, Redmond, Wash.

Appendix 1

Equivalence testing for independent binary endpoints via the Hauck-Anderson corrected classical procedure from Tu[12]

Hypothesis testing approach:

$H_{01}: P_T - P_R \le \theta_1$ vs $H_{11}: P_T - P_R > \theta_1$
$H_{02}: P_T - P_R \ge \theta_2$ vs $H_{12}: P_T - P_R < \theta_2$

where P_T and P_R are the proportion of positive outcomes in the test and reference proportions, respectively, and θ_1 and θ_2 are the lower and upper prespecified limits, respectively. Reject H_{01} and H_{02} when

$T_1 \ge z_\alpha$ and $T_2 \le -z_\alpha$

where

$$T_1 = \frac{(P_T - P_R) - \theta_1 - C}{\sigma}, T_2 = \frac{(P_T - P_R) - \theta_2 + C}{\sigma}$$

and

$$\sigma = \sqrt{\frac{P_T(1 - P_T)}{n_T - 1} + \frac{P_R(1 - P_R)}{n_R - 1}} \text{ and } C = \frac{1}{2 \min(n_T, n_R)}$$

where n_T and n_R are the number of animals in the treatment and reference groups, respectively, and C is a correction factor, calculated using the sample size of the smaller of the two groups

Confidence limits approach:

$\theta_1 = P_T - P_R - z_\alpha\sigma - C$ and $\theta_2 = P_T - P_R + z_\alpha\sigma + C$

Appendix 2

Equivalence testing for normally distributed independent data[5]

Hypothesis testing approach:

$H_0: \mu_T - \mu_R \le \delta_1$ or $\mu_T - \mu_R \ge \delta_2$ vs $H_1: \delta_1 < \mu_T - \mu_R < \delta_2$

where μ_T and μ_R are the means of the test and reference groups, respectively, and δ_1 and δ_2 are the lower and upper limits of agreement, respectively. If $\delta_i = f_i\mu_R$ and $\theta_i = 1 + f_i$, $i = 1, 2$, the hypotheses can be rewritten as two one-sided equations

$$H_{01}: \frac{\mu_T}{\mu_R} \le \theta_1 \text{ vs } H_{11}: \frac{\mu_T}{\mu_R} > \theta_1$$

$$H_{02}: \frac{\mu_T}{\mu_R} \ge \theta_2 \text{ vs } H_{12}: \frac{\mu_T}{\mu_R} < \theta_2$$

where θ_1 and θ_2 are the lower and upper limits of the equivalence interval for the ratio of μ_T and μ_R. Reject H_{01} and H_{02} when

$T_1 \ge t_{\alpha, n_1 + n_2 - 2}$ and $T_2 \le -t_{\alpha, n_1 + n_2 - 2}$

where

$$T_i = \frac{\overline{X}_T - \theta_i\overline{X}_R}{S_p\sqrt{\left(\frac{1}{n_1} + \frac{\theta_i^2}{n_2}\right)}}, i = 1, 2$$

where $\overline{X}_T$ and $\overline{X}_R$ are means of the treatment and reference groups, n_1 and n_2 are the number of animals in the treatment and reference groups, respectively, and S_p is the pooled SD

Fieller's confidence interval approach:

$$\theta_i = \frac{\overline{X}_T\overline{X}_R \pm \sqrt{(a_R\overline{X}_T^2 + a_T\overline{X}_R^2 - a_Ta_R)}}{\overline{X}^2_R - a_R}, i = 1, 2$$

where

$$a_T = \frac{S_p^2}{n_1}t^2_{\alpha, n_1 + n_2 - 2} \text{ and } a_R = \frac{S_p^2}{n_2}t^2_{\alpha, n_1 + n_2 - 2}$$

Note: Conclude equivalence only if the interval (θ_1, θ_2) is within acceptable limits and $\overline{X}_R^2 > a_R$

References

1. Altman DG, Bland MJ. Absence of evidence is not evidence of absence. *Br Med J* 1995;311:485.
2. Greene WL, Concato J, Feinstein AR. Claims of equivalence in medical research: are they supported by the evidence? *Ann Intern Med* 2000;132:715–722.
3. Farrington CF, Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Stat Med* 1990;9: 1447–1454.
4. Chan ISF. Exact tests of equivalence and efficacy with a non-zero lower bounds for comparative studies. *Stat Med* 1998;17: 1403–1413.
5. Hauschke D, Kieser M, Diletti E, et al. Sample size determination for proving equivalence based on the ratio of two means for normally distributed data. *Stat Med* 1999;18:93–105.
6. Bristol DR. Clinical equivalence. *J Biopharm Stat* 1999;9: 549–561.
7. Pidgen AW. Statistical aspects of bioequivalence—a review. *Xenobiotica* 1992;22:881–893.
8. Bland MJ, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;8:135–160.
9. Gould AL. A practical approach to evaluating population and individual bioequivalence. *Stat Med* 2000;19:2721–2740.
10. Obuchowski NA. Can electronic medical images replace hard-copy film? Defining and testing the equivalence of diagnostic tests. *Stat Med* 2001;20:2845–2863.
11. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986:307–310.
12. Tu D. A comparative study of some statistical procedures in establishing therapeutic equivalence of nonsystemic drugs with binary endpoints. *Drug Inf J* 1997;31:1291–1300.
13. Welsh EM, Gettinby G, Nolan AM. Comparison of a visual analogue scale and a numerical rating scale for assessment of lameness, using sheep as a model. *Am J Vet Res* 1993;54:976–983.
14. Bienzle D, Stanton JB, Embry JM, et al. Evaluation of an in-house centrifugal hematology analyser for use in veterinary practice. *J Am Vet Med Assoc* 2000;217:1195–1200.
15. Wess G, Reusch C. Evaluation of five portable blood glucose meters for use in dogs. *J Am Vet Med Assoc* 2000;216: 203–209.
16. Bland MJ, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet* 1995;346:1085–1087.
17. Tu D. Two one-sided tests procedures in establishing therapeutic equivalence with binary clinical endpoints: fixed sample performances and sample size determination. *J Stat Comput Simul* 1997; 59:271–290.
18. Tu D. On the use of the ratio or the odds ratio of cure rates in therapeutic equivalence clinical trials with binary endpoints. *J Biopharm Stat* 1998;8:263–282.
19. Hauck WW, Anderson S. A comparison of large-sample confidence interval methods for the difference of two binomial probabilities. *Am Stat* 1986;40:318–322.
20. Dunnett CW, Gent M. Significance testing to establish equivalence between treatments, with special reference to data in the form of 2x2 tables. *Biometrics* 1977;33:593–602.
21. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *J Pharmacokinet Biopharm* 1987;15: 657–680.
22. Howe LM, Slater MR, Boothe HW, et al. Long-term outcome of gonadectomy performed at an early age or traditional age in cats. *J Am Vet Med Assoc* 2000;217:1661–1665.
23. Miller DD, Staats SR, Partlo C, et al. Factors associated with the decision to surrender a pet to an animal shelter. *J Am Vet Med Assoc* 1996;209:738–742.
24. Ko JCH, Miyabiyashi T, Mandsager RE, et al. Renal effects of carprofen administered to healthy dogs anesthetized with propofol and isoflurane. *J Am Vet Med Assoc* 2000;217:346–349.
25. Dawson-Saunders D, Trapp RG. *Basic and clinical biostatistics*. 3rd ed. New York: McGraw-Hill Book Co, 2001;134–135.