Type II error and statistical power in reports of small animal clinical trials

Michelle A. Giuffrida, VMD

Objective—To describe reporting of key methodological elements associated with type II error in published reports of small animal randomized controlled trials (RCTs) and to determine the statistical power in a subset of RCTs with negative results.

Design—Descriptive literature survey.


Procedures—Predefined criteria were used to identify trial primary outcomes and classify results as negative or positive. Details of sample size determination and use of confidence intervals in results reporting were recorded. For each 2-group RCT with negative results, the statistical power to detect 25% and 50% relative differences in outcome was calculated.

Results—Of 238 RCTs, 42 (18%) stated a primary outcome, 52 (22%) reported a sample size calculation, and 18 (9%) included a confidence interval around the observed treatment effect. Reports of only 2 (0.8%) RCTs included all 3 elements. Among 103 two-group RCTs with negative results, only 14 (14%) and 40 (39%) were sufficiently powered (β < 0.20) to detect 25% and 50% relative differences in outcome between treatments, respectively.

Conclusions and Clinical Relevance—The present survey found that small animal RCTs with negative results were often underpowered to detect moderate-to-large effect sizes between study groups. Information needed for critical appraisal was missing from most reports. The potential for clinicians to base treatment decisions on inappropriate interpretations of RCTs was worrisome. Design and reporting of small animal RCTs must be improved. (J Am Vet Med Assoc 2014;244:1075–1080)

A well-planned RCT is generally accepted as the best study design to assess comparative treatment efficacy.1,2 Despite internal safeguards against systematic error, RCTs are subject to imprecise results owing to random error, particularly when sample sizes are small.3 Two types of random error can occur (Appendix). When a trial finds no difference in outcome between treatment groups, clinicians must be able to assess the likelihood that an important difference was missed because of type II error.

The probability of making a type II error is represented by β but is more commonly expressed in terms of power (1 – β). Power indicates the probability of obtaining a certain set of statistical results, given a certain sample size.1 When planning a trial, investigators should prespecify the primary study outcome and the minimum difference in outcome (treatment effect) between groups they consider clinically relevant.3 A calculation can be performed to determine the sample size required to have a reasonably high power of detecting the smallest relevant treatment effect, if it exists. By convention, studies with < 80% power (β > 0.20) are considered to have excessively high risk of false-negative results due to type II error.1,3 For a given treatment effect, sample size and power are directly related: a larger sample size will have greater power to detect the effect, and vice versa. Similarly, when power is set (eg, at 80%), the required sample size gets larger as the treatment effect of interest gets smaller.

From the Department of Clinical Studies-Philadelphia, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA 19104. There are no funding sources or conflicts of interest to acknowledge as pertain to this manuscript. Address correspondence to Dr. Giuffrida (michellg@vet.upenn.edu).

ABBREVIATIONS

CI Confidence interval
CONSORT Consolidated Standards of Reporting Trials
RCT Randomized controlled trial

In veterinary medicine, clinical trials involving small animals often have small sample sizes and methodological shortcomings that indicate inadequate planning.6–8 Consequently, many RCTs could have low power from the outset to detect important treatment differences. If an underpowered study finds no significant difference between treatment groups, the results are inconclusive because an important treatment effect has not been ruled out. When a trial with a negative result fails to report power or a measure of precision (eg, CI) around the observed treatment effect, the likelihood of type II error is unknown and the results are not interpretable.4 To clinicians with only a basic understanding of biostatistics, RCTs at risk of type II error will still be considered definitive evidence based on randomized design alone. The relative scarcity of small animal RCTs heightens the concern that unsubstantiated interpretations of trials with negative results could go unchallenged for years, during which time many animals could be denied potentially beneficial treatments.

Reviews of published reports of human-subject RCTs indicate that primary outcome, power, and sample size determination are inconsistently reported, and trials with negative results often lack the power to detect large treatment effects.9–13 These disturbing trends could reasonably also exist in the small animal literature but have not been reported. The purpose of the study reported here was to describe the reporting of key methodological elements associated with type II error in
reports of published small animal clinical RCTs and to determine level of statistical power in a subset with negative results.

Materials and Methods

Sample—The investigator (MAG) identified eligible RCTs through a by-hand search of all items published from 2005 through 2012 in the following veterinary journals: *American Journal of Veterinary Research, Journal of the American Animal Hospital Association, Journal of the American Veterinary Medical Association, Journal of Small Animal Practice, Journal of Veterinary Emergency and Critical Care, Journal of Veterinary Internal Medicine, Journal of Veterinary Pharmacology and Therapeutics, Veterinary and Comparative Oncology, Veterinary Dermatology, Veterinary Record,* and *Veterinary Surgery.* Inclusion criteria were the following methodological features: client-owned canine or feline study subjects, prospective data collection, and randomization of animals to parallel treatment and control groups (placebo or active). Trials with purpose-bred, random-source, and shelter- or colony-housed animals were excluded, as were trials involving historical or healthy controls, crossover, n-of-1, or other nonparallel designs and articles describing the results of multiple trials within a single report.

Data collection—The following data were collected from each RCT: species, sample size, number of study groups, whether a primary outcome measure was clearly specified, whether the trial results were negative or positive, whether an a priori sample size or power calculation was performed and any reported elements used in the calculation, and whether CIs were provided around the observed treatment effects. Because it was anticipated that these data would not be explicitly reported in all studies,4,6,12 the investigator used prespecified decision rules to classify each study’s primary outcome and direction of results. The criteria closely followed those used by Freiman et al12 and Moher et al11 in related human-subject RCT research. An RCT’s primary outcome measure was considered clearly specified if 1 of the following 3 criteria was satisfied: only one outcome measure was reported, the primary outcome was explicitly stated, or the specific outcome used to calculate the study’s sample size or power was reported, in which case this was considered the primary outcome. Randomized controlled trials that did not satisfy any of these criteria were considered not to have a clearly specified primary outcome measure. An RCT was classified as having negative results if 1 of the following 3 criteria was satisfied: the authors explicitly stated that negative or equivalent results were obtained; the authors did not state the results were negative, but the study’s primary outcome measure was clearly specified per criteria and a significant (P < 0.05) difference between groups was not reported for this outcome; or the authors did not state the results were negative, the study’s primary outcome measure was not clearly specified per criteria, multiple outcomes were evaluated, and a significant (P < 0.05) difference between groups was reported for < 50% of outcomes. Results of RCTs that did not satisfy any of these criteria were classified as positive. For trials classified as having negative results, any statements considering whether the study might have missed a clinically relevant treatment effect (beyond simply mentioning that the study size was small) were recorded.

A subset was identified that included reports of all 2-group RCTs with negative results that had continuous, dichotomous, or time-to-event primary outcomes and sufficient statistical information for power analysis. The elements necessary to calculate study power for the primary outcome were extracted from each trial in the subset: number of experimental subjects, ratio of control to experimental subjects, and observed group proportions; group means and control group SD; or group median times to event and recruitment and follow-up intervals (for studies with dichotomous, continuous, and time-to-event outcomes, respectively). If SE was reported in lieu of SD, SD was calculated by multiplying the error term by the square root of the group sample size.13 If neither SD nor SE were reported, the SD was estimated by dividing the range of values by 6.12 For studies with multiple outcomes but no specified primary outcome, best judgment was used to designate a primary outcome from among relevant outcomes pertaining directly to the interventions compared. Information in the title, abstract, introduction, and discussion sections was used to infer the authors’ intentions. When multiple potential primary outcomes were identified, the outcome considered most serious was chosen for power analysis per the method of Moher et al.11

Statistical analysis—Statistical analysis was performed with computer software.13 Descriptive statistics were calculated. Continuous variables (sample size and number of study groups) were not normally distributed on the basis of quantile-quantile plots and were expressed as medians and ranges. Categorical data were expressed as frequencies and percentages.

Power calculations were performed in a separate computer software program.13,14 Two calculations of statistical power were performed for each RCT in the subset of 2-group RCTs with negative results: the power of the study to identify a 25% and a 50% relative difference between groups, on the basis of results for the primary outcomes reported. All power calculations were performed as 2-tailed t tests, Z tests, or log-rank tests as appropriate to the scale of primary outcome measurement. A value of α = 0.05 was used for all calculations.

Results

During the 8 years of review, 240 reports of parallel-group clinical small animal RCTs were identified in the 11 journals (Table 1). Overall, 200 (84%) trials studied dogs and the re...
remainder studied cats. The median number of study subjects was 40 (range, 10 to 445), and the median number of parallel study groups was 2 (range, 2 to 6).

Of the 240 trials, results were classified as negative in 165 (69%) and positive in 73 (30%); 2 (1%) were excluded from further analysis because it was not possible to determine the direction of results. Among trials with negative results, 103 were identified based on explicit statements in the text that results were negative and 62 were classified according to other criteria: in 11 trials, the specified primary outcome was negative, and in 51 trials, a primary outcome was not specified but >50% of outcomes were negative.

A primary outcome measure was clearly specified in 77 (32%) trials: it was explicitly stated in 42 trials, was derived from the sample size calculation in 32 trials, and was the only outcome reported in 3 trials. A priori sample size or power calculation was reported in 52 (22%) trials. Twenty-eight of the 52 trials described the treatment effect on which the sample size was based, and 23 provided all the elements required to replicate the calculation. A CI around at least 1 treatment effect was reported in 22 (9%) trials. Trials with positive results and trials with negative results both enrolled few subjects and were unlikely to specify a primary outcome, report a sample size calculation, or include CIs around the treatment effects (Table 2). Reports of only 2 of 238 (0.8%) trials, one with positive results and the other with negative results, explicitly stated all 3 methodological elements.

Of the 165 trials with negative results, 36 studied multiple treatment groups and reports of 26 included insufficient statistical information for power analysis. Therefore, 103 trials were included in the subset of simple 2-group trials used for power analysis. Only 14 (14%) and 39 (40%) of 103 trials had at least 80% power to detect 25% and 50% relative differences, respectively, between study groups (Figure 1). The median power of all 103 trials to detect a 25% effect size was 20% (range, 0% to 100%), and the median power to detect a 50% effect size was 62% (range, 0% to 100%). Only 14 of the 103 trials made any statement considering the possibility that the study missed a clinically relevant treatment effect.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Negative (n = 165)</th>
<th>Positive (n = 73)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canine</td>
<td>137 (83)</td>
<td>63 (86)</td>
</tr>
<tr>
<td>Sample size</td>
<td>34 (24–64)</td>
<td>51 (30–90)</td>
</tr>
<tr>
<td>No. of parallel groups</td>
<td>2 (2–2)</td>
<td>2 (2–3)</td>
</tr>
<tr>
<td>Primary outcome specified</td>
<td>41 (25)</td>
<td>36 (49)</td>
</tr>
<tr>
<td>Power or sample size calculation</td>
<td>28 (17)</td>
<td>24 (33)</td>
</tr>
<tr>
<td>CI around treatment effect</td>
<td>9 (5)</td>
<td>13 (18)</td>
</tr>
</tbody>
</table>

Data are expressed as number (percentage) or median (interquartile range).

Figure 1—The power of 103 small animal RCTs with negative results to identify 25% (A) and 50% (B) relative differences in primary outcome between treatment groups as significant (P = 0.05). Randomized controlled trials to the left of the dotted lines are underpowered according to conventional standards.

Figure 2—Illustration of post hoc power. Curves represent distributions for control (left) and experimental (right) outcomes. A—The thin vertical line represents the conventional values for type I (α = 0.05) and type II (1 – β = 0.80) error. B—In a post hoc power calculation, the critical value is shifted to correspond to the mean effect size (thick vertical line), so the power to detect it is always 50%. H0 = Null hypothesis. (Adapted from Dyba T, Kampenes VB, Sjoberg DIK. A systematic review of statistical power in software engineering experiments. Inf Softw Technol 2006;48:745–755. Reprinted with permission.)
Discussion

This study documented 2 main issues. The first was that small animal trials with negative results were frequently underpowered to detect moderate to large relative differences in outcome between treatment groups. The second issue was of perhaps even greater concern: readers would not be able to easily identify these underpowered studies because critical information was missing from the published manuscripts. These findings suggest that many RCTs with apparently negative results are in fact inconclusive or cannot be meaningfully interpreted. Inappropriate changes in patient management could occur if readers misinterpret the results of these trials.

Most of the trials with negative results in this review had sample sizes that were too small to reliably identify a 50% better outcome in one group relative to the other, according to conventional standards of significance. Therefore, the potential for these studies to miss a clinically relevant treatment effect due to type II error is quite high. It is not clear that authors recognized this problem, given how few reports contained any statement considering whether the study could have missed a meaningful effect. Treatment effects > 50% are no doubt clinically relevant; however, in most instances, it would be unreasonable to expect them to exist.2 Treatments with very large therapeutic effects are likely to be identified through clinical observation or other less stringent methodologies.10 Clinical trials are more often used to test treatments of controversial efficacy, which cannot be expected to exert more than moderate treatment effects.16 Some authors have argued that underpowered trials are still valuable because they can be combined in systematic reviews or meta-analyses.17 However, this is only true if a study is otherwise free of major biases and key methodological elements are clearly reported, which was generally not the case for these trials nor for those analyzed in prior reviews.6,9,18,19 Other authors raise concerns that underpowered trials are unethical because they monopolize resources, inconvenience the study participants, and deprive patients of potentially beneficial treatments but do not meaningfully inform clinical practice.9,20–22

Investigators can avoid these ethical quandaries by adhering to recommended standards of study design and reporting, such as those outlined in the CONSORT statement. The CONSORT statement recommends that all trials explicitly define a primary outcome and describe how the sample size was determined (including all elements used in calculations).3 An a priori power or sample size calculation indicates that investigators purposefully considered the primary outcome and the size of the effect that would be clinically important.22,23 It also provides a safeguard so that investigators do not abandon the original hypothesis and shift their attention to questions whose answers appear most favorable once the data are obtained.23,24 Even so, reports of several studies included a sample size calculation but then failed to report the results of the outcome on which the calculation was based. Others indicated that a calculation had been performed but failed to provide any details such as what outcome or treatment effect it was based on. In many situations, no extensive prior data are available on which to base a power calculation; however, investigators should still be concerned with detecting a clinically relevant effect and should explain how the sample size was determined with this in mind.

The observation that most reports of studies did not state a primary outcome measure or sample size calculation is consistent with other literature surveys of veterinary trials. Lund et al6 reviewed the reporting of methodological features in reports of 23 small animal RCTs published between 1989 and 1990 and found that none reported the power of the study or how the sample size was selected. A more recent review by Sargeant et al7 assessed a sample of reports of small animal controlled trials published from 2006 to 2008 for reporting of items included in the CONSORT checklist. Among 85 trials, a primary outcome was reported for only 6 (7%) and a sample size calculation was reported for only 1 (1%). The higher rates of primary outcome and sample size reporting in the study reported here could reflect improved awareness of methodology or reporting standards on the part of veterinary authors. Nevertheless, most reports of RCTs still failed to include these important details. Compared with trials with positive results, RCTs with negative results in this study had smaller sample sizes and were less likely to have a primary outcome or sample size calculation reported, highlighting the potential for false-negative results due to inadequate study design. In reviews of human subject RCTs, larger sample size is both an independent predictor of positive study outcomes25 and is associated with higher overall methodological quality.26 The authors of these studies also postulate that negative findings are often due to inadequate power stemming from deficient study design. The persistent absence of sample size and power calculations in reports of veterinary RCTs is difficult to explain. Particularly perplexing is the fact that many of these trials were published in veterinary journals that openly endorse the CONSORT statement and related veterinary reporting guidelines, which unequivocally recommend that RCTs report a primary outcome and how the sample size was determined. It is likely that editors will need to demand adherence to relevant reporting guidelines as a prerequisite for publication, if any meaningful change in study design and reporting is to occur.27

Once the results of a study are known, the pretrial power can be used to help interpret the potential for type II error in a result that does not meet the criteria for significance. The CONSORT group also recommends that authors report the numeric outcome for each treatment group as well as a measure of the contrast between groups (eg, the treatment effect, expressed as a difference in means, and OR or hazard ratio) with a surrounding CI.3 This approach provides more information than the arbitrary dichotomy of significance because it identifies a range of possible treatment effects supported by the data.28,29 If the CI includes an effect size that would be considered clinically relevant, the result is considered inconclusive rather than negative. A post hoc power calculation of a negative result will always find that the study had low power to detect the observed difference as significant,22,23 and is not recommended.3,13 This occurs because the value of β associated with the mean observed effect size will always be 0.50, or 50% power (Figure 2).30 Confidence intervals
were rarely reported around the effect size among the RCTs in this review, although not all trials reported effect sizes, and several trials included CIs around the individual group results instead. Confidence intervals are confusing for many researchers, which could explain their inappropriate use or absence. However, reporting a measure of precision around the observed treatment effect greatly facilitates meaningful interpretation of the trial, particularly among those with negative results.

Many authors of trials with negative results noted that their studies’ small sample sizes were a limitation on the results, without expanding further on why this should be so. A few authors misleadingly stated that if the sample size had been larger, the significant result they desired to identify would become evident. Others justified an underpowered study by commenting that the sample size necessary to identify the authors’ desired result would be prohibitively large. These statements indicate that authors often have an overly simplistic view of the relationship between sample size and desired results. Large sample size is not the only way to increase study power. The study hypothesis, the effect size of interest, the format of the outcome data (eg, dichotomous vs continuous), and the statistical tests used are just a few examples of design and analysis choices that have considerable impact on a study’s power. Furthermore, the optimal choice of type I and type II errors is not necessarily the conventional values of \( \alpha = 0.05 \) and \( \beta = 0.20 \) but varies according to the available sample sizes and plausible effect sizes expected in different fields. Thus, it would be extremely useful for veterinary investigators to enlist the aid of biostatisticians at the outset of any proposed trial. This collaboration could improve the ability of investigators to use methods other than large sample size to design studies that plausibly address clinical questions. It would also shift the burden of statistical decision making and explanation toward those with the appropriate expertise.

Articles included in this study were published in a select number of English-language veterinary journals, so the results cannot necessarily be generalized to all small animal RCTs. Decision rules were used to identify primary outcomes and establish the direction of results because authors failed to adequately convey their intentions in the published manuscripts. Consequently, there were some discrepancies between how the trials were classified in this study and how the authors appeared to interpret the results. Although results of 165 trials were classified as negative, the authors did not always acknowledge that the results were negative. Many authors appeared to conclude a positive treatment effect despite a clearly specified primary outcome that was negative, mostly negative outcomes with a few significant but clinically irrelevant differences, or absolutely no outcome differences between groups. If each trial had specified a primary outcome, it is almost certain that some trials would have been classified differently with respect to that outcome, the direction of results, or the study power. This is an apt illustration of the interpretive obstacles created by incomplete study planning and reporting. The inability of small animal RCTs to detect clinically relevant effects or adequately convey trial methodology and results to the veterinary community should be concerning to both investigators and clinicians. Veterinary researchers could improve the study design and analysis of their RCTs by consulting with experienced clinical trials biostatisticians and epidemiologists during the initial development of any proposed trial. Methodological resources such as the CONSORT statement, which are intended to provide direction for both the design and reporting aspects of trials, should be consulted as early as possible. Careful planning can prevent animals from being enrolled in trials whose results cannot be interpreted or that have no chance at the outset of detecting relevant treatment effects. Authors preparing manuscripts of small animal RCTs should follow the CONSORT checklist to ensure that methodological details and results are reported in a clear and complete manner. Transparent reporting enables readers to assess the validity of study findings and provides a means for researchers to replicate a trial’s methods or extract data for systematic reviews and meta-analyses. Readers of reports of veterinary RCTs should be aware of the prevalence of underpowered studies and realize that author interpretations of results could be incorrect or misleading. When seeking the best available evidence to guide clinical decision making, veterinarians must critically appraise published manuscripts for details of design, conduct, and analysis before applying the results to their patients.

References

Efficacy of a flexible schedule for administration of a *Leptospira borgpetersenii* serovar Hardjo bacterin to beef calves

Victor S. Cortese et al

**Objective**—To determine whether a flexible vaccination regimen provides beef calves protection against challenge exposure with a virulent *Leptospira borgpetersenii* serovar Hardjo isolate.

**Animals**—Fifty-five 4-week-old calves seronegative for *L. borgpetersenii* serovar Hardjo.

**Procedures**—Calves were assigned to 3 groups and administered 2 doses of adjuvant (control calves; n = 11), 1 dose of serovar Hardjo bacterin and 1 dose of adjuvant (22), or 2 doses of the serovar Hardjo bacterin (22); there was a 16-week interval between dose administrations. Three weeks after the second dose, all calves were challenged by conjunctival instillation of a heterologous strain of *L. borgpetersenii* serovar Hardjo for 3 consecutive days. Urine samples for leptospiral culture were collected for 5 weeks after challenge exposure; at that time, all calves were euthanized and kidney samples collected for leptospiral culture.

**Results**—Antibody titers increased in both leptospiral-vaccinated groups of calves. A significant increase in antibody titers against *L. borgpetersenii* serovar Hardjo was detected after administration of the second dose of *L. borgpetersenii* serovar Hardjo bacterin and challenge exposure. In 10 of 11 adjuvant-treated control calves, serovar Hardjo was isolated from both urine and kidney samples. *Leptospira borgpetersenii* serovar Hardjo was not isolated from the urine or kidney samples obtained from any of the 21 remaining calves that received 1 dose of bacterin or the 20 remaining calves that received 2 doses of bacterin.

**Conclusions and Clinical Relevance**—Protection in young calves was induced by vaccination with 1 or 2 doses of a serovar Hardjo bacterin. (Am J Vet Res 2014;75:507–512)