Journal of Veterinary Internal Medicine
Open Access
ACVIM American College of Veterinary Internal Medicine

*Perspective*
*J Vet Intern Med* 2017

# Clinical Trials in Veterinary Medicine: A New Era Brings New Challenges

M.A. Oyama, S.S. Ellenberg, and P.A. Shaw

Randomized clinical trials (RCTs) are among the most rigorous ways to determine the causal relationship between an intervention and important clinical outcome. Their use in veterinary medicine has become increasingly common, and as is often the case, with progress comes new challenges. Randomized clinical trials yield important answers, but results from these studies can be unhelpful or even misleading unless the study design and reporting are carried out with care. Herein, we offer some perspective on several emerging challenges associated with RCTs, including use of composite endpoints, the reporting of different forms of risk, analysis in the presence of missing data, and issues of reporting and safety assessment. These topics are explored in the context of previously reported veterinary internal medicine studies as well as through illustrative examples with hypothetical data sets. Moreover, many insights germane to RCTs in veterinary internal medicine can be drawn from the wealth of experience with RCTs in the human medical field. A better understanding of the issues presented here can help improve the design, interpretation, and reporting of veterinary RCTs.

**Key words:** Competing risk; Endpoints; Epidemiology; Study design.

A glance through recent issues of the Journal of Veterinary Internal Medicine identifies prospective randomized clinical trials (RCTs) involving dogs,[1–9] cats,[10–13] cows,[14,15] and horses[16,17] that cover the fields of neurology,[1,8] oncology,[3,5] cardiology,[4,6,12] and internal medicine.[2,9–11,13–17] Collectively, these findings suggest that we have entered into an era of prospective veterinary RCTs. Without question, this is an achievement worth celebrating, but as is often the case, with progress comes new challenges. Randomized clinical trials yield important answers, but results from these studies can be difficult to interpret, incomplete, unhelpful, or even misleading unless the study design, execution, reporting, and interpretation are carried out with care. The methodology of well-designed and reported veterinary RCTs has been the subject of several previous publications.[18–21] In a very simple sense, the best RCTs foster better everyday clinical decisions made at the patient's side. This review addresses emerging challenges associated with RCTs, including use of composite endpoints; the reporting of different forms of risk; including baseline, relative, and absolute risk; analysis in the presence of missing data; and, issues of reporting and safety assessment. These topics are explored in the context of previously reported internal medicine studies as well as through illustrative examples with hypothetical data sets. Additionally, many insights can be drawn from the wealth of experience with RCTs in the human medical field. A better understanding of the issues presented here can help improve the design, interpretation, and reporting of veterinary RCTs.

**Abbreviations:**

| | |
|---|---|
| CHF | congestive heart failure |
| DSMB | data and safety monitoring board |
| HR | hazard ratio |
| ITT | intention-to-treat |
| PP | per-protocol |
| RCTs | randomized clinical trials |
| TTE | time-to-event |

## Composite Endpoints

A critical choice when designing RCTs is the choice of the study outcomes or endpoints. The primary endpoint is the main event that the treatment being evaluated is intended to beneficially affect. Additional, or secondary, endpoints also are commonly examined to gather related evidence that supports the primary outcome. Endpoints and any expected treatment benefit must be sufficiently important to the animal and owner with respect to morbidity, mortality, quality of life, or some combination of these. Many endpoints involve a dichotomous (i.e, yes versus no) event such as death, tumor relapse, or hospitalization for worsening disease. In RCTs in human medicine,[22] and increasingly in veterinary RCTs,[4,6,23–26] multiple clinical endpoints are combined into a single "composite" endpoint, defined as the occurrence or time to first of any of the component outcomes, or more simply, time-to-event (TTE). Ideally, the selected components should all be related to the primary outcome, expected to respond to the treatment under study, and occur with similar frequency.[27,28]

*From the Department of Clinical Studies-Philadelphia, School of Veterinary Medicine, (Oyama); Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, (Oyama, Ellenberg, Shaw); and Division of Biostatistics, Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA (Ellenberg, Shaw).*

*Work was done at the University of Pennsylvania.*

*Work not presented at any meetings.*

*Corresponding author: M.A. Oyama, Department of Clinical Studies-Philadelphia, School of Veterinary Medicine, University of Pennsylvania, 3900 Delancey Street, Philadelphia, PA 19104; e-mail: maoyama@upenn.edu*

*Submitted February 27, 2017; Revised March 17, 2017; Accepted April 20, 2017.*



*DOI: 10.1111/jvim.14744*

For example, studies in human medicine of coronary artery disease commonly combine several major adverse clinical events into a composite endpoint.[29] Such events typically include cardiovascular death, reinfarction, stroke, or need for coronary vessel revascularization; the TTE analysis takes the timing of whichever event occurs first as the endpoint.[30] An example of a composite endpoint in veterinary medicine is the use of time to either onset of congestive heart failure (CHF) or sudden death in studies of dogs with dilated cardiomyopathy.[6,24] A dog was considered to have experienced the endpoint *either* by developing CHF *or* by dying suddenly, whichever came first. Similarly, in a study[5] evaluating whether an intervention prevented tumor recurrence or death due to mammary cancer in dogs, the time to the first of either event was regarded as an endpoint.

Investigators utilize a composite endpoint primarily for efficiency.[31] For TTE outcomes, the number of events drives study power,[32] which is the probability the study will detect a true underlying difference between treatment and control groups. In these studies, the number of events can be increased by enlarging the sample size or by extending the duration of follow-up. Use of a composite TTE outcome can markedly decrease the sample size (and cost) of a study by increasing the number of expected events.[28,31] For example, consider planning for a hypothetical study to detect a hazard ratio of 0.65 at a significance level of 0.05 with a desired power of 80%. If 50% of subjects in the control group are expected to experience the endpoint, 406 subjects are required (i.e, 203 in the control group and 203 in the treatment group), but if 80% of control subjects reach the endpoint, only 242 subjects (i.e, 121 in each group) are needed. By enriching the number of expected events by use of a composite endpoint, what might have been an unrealistic veterinary study in terms of patient numbers suddenly becomes feasible.

Another advantage of composite endpoints relates to the issue of multiplicity. Having multiple primary endpoints instead of a single composite endpoint increases the probability that 1 of those endpoints will be significantly different between groups merely by chance. For example, if rather than combining 3 separate clinical events into a single composite endpoint, an investigator chooses to evaluate each component separately, the possibility that the treatment comparisons for 1 of these 3 outcomes will be significant merely by chance is 14.3% ($1-[1-0.05]^3$), assuming independence of the endpoints and the typical 0.05 significance level. This problem rapidly grows such that in a study with 14 primary outcomes, the chance of 1 being significant by chance is >50%, again assuming independence. Typically, the outcomes comprising a composite endpoint are correlated to some degree, thereby decreasing the probability of spurious chance findings, but the problem can remain substantial. To counter the problem of multiplicity (i.e, to control the probability that a truly null hypothesis would be declared significant), the threshold for statistical significance for each primary outcome must be made more stringent than the typical threshold of 0.05. This will decrease study power or require more subjects to maintain the desired power for any single endpoint. Use of a composite endpoint, requiring only a single significance test, eliminates the multiplicity problem, but raises other issues, as we will discuss below.

Interpretation of results from RCTs involving a composite endpoint can be illustrated by considering 1 of the aforementioned studies[24] in dogs with dilated cardiomyopathy. The study found that treatment significantly decreased risk of experiencing the composite endpoint by approximately 70% compared to placebo. Yet, when looking at each endpoint separately, neither of the individual components, namely time-to-first onset of CHF or sudden cardiac death, was significant on its own. On the surface, this might seem like a dubious result in which pimobendan prevented neither CHF nor sudden death. The driving motivation behind use of the composite endpoint however was to decrease the necessary number of patients, and thus any evaluation of the individual components would likely be underpowered. In this particular instance, veterinarians should find it reassuring that although neither of the individual components was significant, both trended strongly in the direction that was favorable to the active treatment. Although a trial with a composite endpoint offers less evidence for benefit on the individual endpoints than 1 powered for that individual endpoint, a reasonable inference is that given a larger number of patients, both endpoint components would have continued toward benefit and eventually achieved statistical significance. This is the ideal situation for use of a composite endpoint—the treatment shows consistent benefit on all components, so that the increased power helps to confirm an overall treatment benefit on the composite.

A more difficult scenario arises when the overall composite is positive but ≥1 individual components are decidedly neutral or even trend in the opposite direction. This problem is magnified if there are considerable differences in the clinical severity of the individual components (e.g, if both nonfatal and fatal components are used simultaneously).[22,28,31,33,34] Examples from both the human and veterinary literature help illustrate this point. One study[35] evaluated human patients with diabetes at risk for heart disease with a composite endpoint of nonfatal myocardial infarction or coronary-related death, and found that treatment significantly lowered the risk of nonfatal infarction by 24%, but the rate of coronary-related death tended to increase in patients receiving treatment. Another example from human medicine is a study[36] that reported that treatment with an insulin sensitizer significantly decreased risk of a composite endpoint consisting of new onset of diabetes or death by 60%. An accompanying editorial[28] raised the following important issues: was there a decrease in both diabetes and death and were the 2 outcomes just as likely to occur? It turned out that >94% of the outcomes experienced by the study participants were onset of diabetes, and whereas the relative risk of developing diabetes in patients receiving treatment was significantly decreased at 0.40 (95% confidence interval

[CI], 0.35–0.46), the relative risk of death associated with treatment appeared unchanged at 0.91 (95% CI, 0.55–1.49).[36] Despite the use of the composite endpoint, these examples clearly provide almost no information about the treatment effect on death, and the treatment should not be promoted as preventing mortality.

Next, consider a veterinary study[6] in Irish wolfhound dogs that evaluated the effect of treatment on a composite endpoint of first onset of CHF or sudden cardiac death. Of the 25 dogs that reached the composite endpoint, 21/25 (84%) experienced an episode of CHF, whereas only 4/25 (16%) experienced sudden death. These examples highlight the difficulty in examining individual components, particularly if the number of patients experiencing a particular component is relatively small. The estimate of treatment efficacy for these types of rarely occurring components can be extremely ambiguous. For example, in a study[4] utilizing a composite endpoint of first onset of CHF or cardiovascular death in dogs with preclinical mitral valve disease, a relatively low incidence of cardiovascular death resulted in a treatment-related risk reduction of 4% with extremely wide 95% CI ranging from a 55% decrease in risk to a 210% increase in risk specific to this component. Difficulties such as those described in these examples are particularly common in studies with a mixture of fatal and nonfatal endpoint components. Mortality is obviously an important clinical event, but in trials of particularly short but practical duration, the majority of participants experience the less severe or nonfatal clinical endpoints, which then subsequently drives the analysis.[29,33,37] Over one-third of medical trials in humans utilizing composite endpoints involving mortality components demonstrated a significant overall result that did not include an appreciable contribution by the mortality component.[33] This situation is most problematic if the effect on the most severe but less frequent endpoint, such as mortality, is in the opposite direction of the effect on the remaining components of the composite.

There are no universally accepted solutions to the problems inherent in use of composite endpoints, but a number of approaches have been proposed.[38] One potential way to balance the importance of different components is to weight them according to their severity or whether they are fatal or nonfatal. For instance, in a hypothetical study of patients with neoplasia, an endpoint of death due to metastatic disease could be weighted 3 times greater than an endpoint of surgery for tumor reoccurrence, and need for surgery could be weighted 1.2 times greater than the need for additional chemotherapy. In practice, weighting of different outcomes is a difficult task. Despite methodology that tries to account for patient preference,[39] the weights are undeniably arbitrary. Is death really 3 times as bad as surgery? How much better or worse is surgery than chemotherapy? For these reasons, weighting has not been widely used. Another strategy is the careful selection of secondary endpoints for analysis, such as all-cause mortality, for which the overall survival between groups is based on death for any reason. In studies[4] with low rates of cause-specific mortality, the finding that patients receiving treatment survived longer no matter what the cause of death is reassuring. In summary, use of composite endpoints attempts to strike a balance between feasibility of a trial and results devoid of ambiguity around the individual components. In practice, constructing composite endpoints that meet these criteria can be very challenging.

## Analysis of Patient Populations with Missing Outcome Data

The ability of RCTs to produce an unbiased estimate of effect between 2 treatment groups is threatened when outcome data are missing. If a study involving 200 randomized patients ultimately collects outcome data on only 120 individuals, the comparability of the treatment groups is no longer protected by randomization because those with missing data might be systematically different from those who provide data. Missing data can arise from a variety of causes. Investigators might remove randomized individuals from study analysis for reasons such as failure to comply with the study protocol or concurrent use of prohibited medications. Missing data also occurs when patients are lost to follow-up or when the animal was withdrawn early from the study. Investigators tend to ignore the problem of missing data by assuming that withdrawals and losses between groups are due to random chance (i.e, are independent from the treatment or outcome) and can therefore be ignored. Many studies then exclusively analyze the subset of patients that successfully completed the protocol, have outcome data, and complied with the study protocol (i.e, "per-protocol" [PP]), while ignoring that fact that missing data might not be missing completely at random and that treatment comparisons subsequently might be biased.

The only reliable way to produce a truly unbiased estimate in the face of missing data is to perform an "intent-to-treat" (ITT) analysis, in which every subject that was randomized regardless of subject compliance with the protocol is included.[40] An important implication of ITT analysis is that studies should gather outcome data on every randomized study patient, regardless of whether or not the patient fully or properly completed the study.[40] Although this is not always possible, the closer one comes to including outcome data on all subjects, the less concern there will be about biased comparisons.[41] Where outcome data due to withdrawals and losses are missing, these cases still can contribute to the ITT analysis by providing valuable information either up until the time they were withdrawn or lost by methods to compensate for the missing data, such as multiple imputation or inverse probability weighting, use of best/worse case scenarios, or other sensitivity analyses.[40,42,43] Intent-to-treat analysis provides a conservative estimate of effect because of potential dilution from early withdrawals from study treatment and decreases the likelihood of a type I (i.e, false positive) error.[44] Intent-to-treat analysis also tends to mimic the interventions effectiveness in the real

clinical world wherein these types of withdrawals and losses occur.[41]

Similarly, there are certain advantages and disadvantages associated with PP analysis. By excluding any individual that did not wholly and completely adhere to the treatment protocol, the PP analysis, in principle, should closely reflect the treatment effect and the underlying scientific basis for its effect. However, a major disadvantage of the PP analysis is that the reason(s) for the missing data might be related to the intervention or outcome. If so, we might find, for example, a difference between groups that is largely due to removing those with the poorest prognosis from 1 of the groups. A classic RCT that demonstrates the potential for this bias is Coronary Drug Project,[45] which examined the efficacy of clofibrate on survival in human patients at high risk of dying from heart disease. After adjusting for known risk factors, the mortality rate in those with poor adherence to the study protocol was higher regardless of whether they were receiving active treatment or placebo, setting up a situation in which censoring for protocol violations was no longer completely at random. In trials designed to show the superiority of 1 intervention over another or over placebo, the more conservative ITT analysis is usually the primary analysis and PP, if considered, is a secondary analysis.[41] When both the ITT and PP analyses lead to the same conclusion, "the confidence in the trial results is increased, bearing in mind, however, that the need to exclude a substantial proportion of subjects from the [PP] analysis throws some doubt on the overall validity of the trial."[41]

The use and reporting of ITT and PP methods in both the human[42,46,47] and veterinary[3–9,13] medical literature varies widely. In many studies, key pieces of information, including original number of patients recruited and randomized, how missing data was treated, and exactly which analysis methods and data sets were used, are lacking.[48,49] Readers are dependent on such data to make informed decisions about RCT results. In a previous survey[44] of RCTs in humans published from 2001 to 2003, 15% of RCTs failed to reach the same statistical conclusion between the 2 analysis methods. In approximately half of these studies, the ITT analysis achieved significance whereas the PP analysis did not, and in the other half the PP analysis was significant whereas the ITT analysis was not. The amount and transparency of information regarding analysis sets are enhanced when journals and authors follow the CONSORT guidelines for the reporting of RCT results in humans.[47] These guidelines, which currently are required by the majority of PubMed-indexed journals for the reporting of RCTs in humans, specify reporting of the number of subjects screened, randomized and ultimately analyzed, with reasons given for those excluded from analysis.[50] In the veterinary profession, well-reported RCTs[4,7,9] in companion animals increasingly include a structured accounting of patient randomization, number and reasons for withdrawal and loss, and the descriptions of both the patient ITT and PP analysis sets in a so-called CONSORT diagram. Studies involving livestock and food safely issues are encouraged to follow analogous standards specific to these types of studies as set forth by the Reporting Guidelines for Randomized Controlled Trials for Livestock and Food Safety (REFLECT) guidelines.[51]

## Risk: Baseline, Relative, and Absolute

Randomized clinical trials provide a group estimate of effect, but the clinician and animal owner want to make clinical decisions relative to an individual animal. Thus, one wants to assess characteristics of the patient population in which the trial established benefit and whether the found treatment effects apply equally across different patient types and individuals. Important baseline variables can affect a patient's prognosis and whether or not a patient will experience treatment benefit. High baseline risk opens the opportunity for a favorable risk-benefit ratio,[52] whereas the patients with low risk might gain little or no treatment benefit.[53] In the case of treatments that are associated with a risk of adverse effects, the benefit-harm ratio might even be reversed for low-risk patients.[54] In addition to baseline risk, 2 distinct types of risk comparisons should be considered. The first of these is a relative risk comparison. The hazard ratio, a potential summary of relative risk, has been mentioned previously. The hazard is a somewhat esoteric mathematical quantity that assesses the instantaneous risk of having the event in the next small time interval, among those still at risk. The commonly used log-rank test and Cox proportional hazards regression model estimate the hazard ratio between groups and assume this ratio is constant throughout the study.[55,56] A more intuitive measure of relative risk is the risk ratio (i.e, the ratio of probabilities of event occurrence between 2 groups at a given time [p1/p2]).[56]

The second type of risk comparison involves examining the absolute difference in the probability of event occurrence between groups of individuals over a specified amount of time (i.e, p1-p2).[57] The distinction between these different forms of risk assessment, relative versus absolute, is important because individual treatment benefit is critically dependent not only on the relative risk, but also on the magnitude of the baseline risk and the *absolute risk reduction* anticipated from treatment.[58] Consider, for example, a hypothetical treatment for cancer that is associated with a 15% relative risk reduction in mortality compared to no treatment that is uniform across all patient subgroups (Table 1). Consider next that patients are stratified by their baseline risk of mortality at the outset of the study into low- and high-risk groups. Such an exercise indicates that the overall absolute treatment benefit is driven almost entirely by the subgroup of patients at highest risk. The absolute risk reduction in patients with low risk at baseline is exceedingly small simply because their risk of dying was low at the outset. Another way to consider absolute risk reduction is to calculate the number of patients needing to undergo treatment in order to have 1 patient benefit. Numerically, the number needed to treat is the inverse of the absolute risk

**Table 1.** Comparison of the relative risk reduction (RRR), absolute risk reduction (ARR) and number needed to treat (NNT) in a hypothetical study that reduces risk of death by 15% in patients receiving treatment.

| | Control Death Rate (%) | Treatment Death Rate (%) | RRR | ARR (%) | NNT |
|---|---|---|---|---|---|
| High-risk patients | 50 | 42.5 | 0.85 | 7.5 | 13 |
| Low-risk patients | 10 | 8.5 | 0.85 | 1.5 | 67 |

Patients have been stratified into those with low and high baseline risk for death at the study outset. While treatment is associated with a lower risk for death in both groups, the ARR for death (i.e, the difference between the control and treatment death rates) in the low-risk patients is extremely small, primarily because these patients were at low risk for death to begin with. NNT is the inverse of the ARR and represents the number of patients needing to be treated in order for 1 patient to gain benefit and is substantially higher in the low- versus high-risk group. Moreover, if the hypothetical treatment happens to be associated with adverse effects in more than 1.5% of patients treated, the net absolute effect of treatment might be harm to the low-risk patient group.

reduction.[57] Thus, in the example of this study, most animals in the low-risk group (66/67; 98.5%) are expected to gain no benefit from treatment for every 1 dog that does. Moreover, if the treatment has even a small risk for serious adverse effects, in this case anything >1/67 (1.5%), treating the low-risk group actually could cause more net harm than benefit.

In human medicine, baseline risk scores based on specific demographic and disease characteristics can be derived from large observational data sets. For instance, the Gail Breast Cancer Risk Assessment Score was developed from a database of nearly 6,000 women[59] and estimates the risk of developing breast cancer over specific periods of time. The score includes influential variables such as family history, race and ethnicity, presence or absence of certain genetic mutations, and age. Another example is the Framingham Heart Study, which utilized 12-year follow-up of >8,400 individuals to predict future cardiovascular disease based on age, smoking status, cholesterol, presence or absence of diabetes, and blood pressure.[60] Predictive data such as these are increasingly available in veterinary patients across a variety of disease conditions, including neoplasia,[61–63] cardiovascular disease,[64,65] and various diseases in horses[66,67] and cows.[68] Incorporation of baseline risk in RCTs potentially can help estimate treatment effects in heterogeneous patient populations with greater accuracy and help clinicians select individuals that are most likely to experience treatment benefit.

In the previous example of low- and high-risk patients, we assumed that the decrease in relative risk would be uniform across the entire cohort of patients regardless of baseline risk. Different patient subgroups however may experience different levels of relative risk reduction. In principle, this "heterogeneity of treatment effect" across subgroups can be statistically evaluated by assessing treatment by subgroup interactions, which involves testing whether the observed differences across subgroups are more than would be expected by chance.[69] In many instances, power to detect interaction effects will be low in trials with sample size calculated for the overall treatment effect.[34,69] This is not to say such analyses should be avoided; exploratory analyses can be valuable in suggesting need for further study, but results of such analyses need to be interpreted very conservatively. In the absence of a prespecified and well-powered interaction effect, the overall treatment effect generally will be the most reliable estimate of efficacy.

To frame the concepts of baseline, relative, and absolute risk in a clinical scenario, consider 2 different hypothetical statements presented to a dog owner who is contemplating adjunctive chemotherapy for a geriatric dog after surgical resection of a tumor. Which statement is more helpful to the owner? The first statement, based solely on relative risk, is as follows: "Chemotherapy will, on average, decrease the risk of tumor recurrence by 50% compared to not giving chemotherapy." This sounds very promising, but only if the risk of tumor recurrence is high, associated adverse effects are tolerable or of low risk, cost of chemotherapy is reasonable, and if the risk of dying in the interim from some unrelated cause is low. Consider next a statement that takes all 3 risk types into consideration as follows: "In dogs with similar baseline risk as your dog, the probability of tumor recurrence in the next 24 months is 8%, and treatment will, on average, decrease this probability to 4%. There is a 3% chance of serious adverse effects from the treatment, and in the interim, there is a 50% chance that your dog will die from its concurrent renal disease rather than the tumor." This second statement provides a much clearer basis for decision-making. If the baseline absolute risk for an event is low, likelihood of an unrelated competing risk event is high, and if there exists even a small risk for treatment-related harm, treatment is unlikely to be beneficial (and in the worse case could actually be harmful) regardless of the reported group effect.[70]

## Other Issues in Clinical Trials in Veterinary Patients

The benefits of well-designed RCTs are self-evident, whereas harm from poorly designed RCTs is more insidious. Poor RCTs waste valuable resources, complicate future research, provide false hope to owners, and potentially could jeopardize patient safety. One of the easiest ways for investigators to insure good design is to carefully predefine their endpoints and statistical plan, preferably with the input of a biostatistician.[21,34,71] The International Committee of Medical Journals Editors (ICMJE) requires registration of clinical trials in a public trials registry at or before the time of first patient enrollment as a condition of consideration for publication.[72,73] For United States journals, this means investigators leading RCTs in humans are required to make their study endpoints and statistical plan public before the start of the study by registering the study on a

publically accessible website.[a] This practice protects the investigator from charges of posthoc data dredging and protects the consumer from being misled by poor analytical practices. The website also tracks the results of trials regardless of whether the end result was confirmatory or null, thereby avoiding the bias that can occur when journals only accept or authors only submit positive trials for publication. Veterinary journals adopting a set of requirements for the publication of RCTs, as put forth by ICMJE, could further strengthen the quality of research. Organizations such as the American Veterinary Medical Association[b] and VetAllTrials consortium[c] recently have launched voluntary public registries for veterinary clinical trials that include description of the trial objectives, potential benefits and risks, and criteria for enrollment.

The number and size of veterinary RCTs relative to the human medical field is small, which presents a tremendous obstacle.[20] A bedrock principle of the scientific endeavor is replicability.[74] The more replicable a finding is, the more likely it is to be valid. If findings from particular RCTs are inconclusive or unexpected, the standard recommendation is to "collect more substantial evidence on the issue,"[34] which usually means additional RCTs. Clinical practice guidelines in human medicine typically rely on confirmatory findings from multiple RCTs, often involving many hundreds or thousands of subjects. Veterinary investigators should seek to "affirm and confirm" findings to the extent that is possible despite the relatively constrained resources available within the veterinary profession. In our opinion, of this study, many veterinarians and animal owners are fully willing to participate in clinical trials, and thus, the limiting factor is 1 of resources and infrastructure. In instances where ethical concerns over repeating a study that showed a new treatment had great benefit or low incidence of disease limit patient enrollment, the veterinary profession might look to studies of cancer or rare or "orphan" human diseases (i.e, those typically affecting <200,000 humans in the United States) for ideas on how to deal with the inability to perform repeated clinical trials.[75] In this respect, the Food and Drug Administration (FDA) offers some general considerations for obtaining "adequate and well-controlled" evidence from single studies and to increase the amount of supporting evidence in the form of consistency across multiple outcomes, statistically persuasive findings, extrapolation from existing studies, or evidence from closely related diseases.[76] The FDA stresses that proper study design and planning are even more critical in these situations than for more common diseases.[76] Aspects of study design that might be particularly suited to relatively small RCTs have been previously reviewed.[77]

Issues involving design of RCTs do not necessarily stop once a trial is underway. Independent data and safety monitoring boards (DSMBs) are utilized in many RCTs in human medicine.[78–80] They comprise clinicians and statisticians not directly involved in the planning or execution of the trial or product being studied. The primary purpose of DSMBs is to oversee patient safety, primarily with respect to adverse events. In many cases, another important component of this oversight is the review of prescheduled interim analyses of unblinded outcome data to determine whether there is sufficient evidence to terminate the study before its scheduled completion of accrual and follow-up, thereby either accelerating the use of superior treatments in the population or decreasing patient exposure to an ineffective or unsafe drug. The analysis of unblinded data necessitates a reviewing body such as the DSMB that is independent of the daily operations of the trial and who can maintain the confidentiality of interim results. If early stopping is considered (and it need not be except in trials where the primary outcome is death or another serious event or those that involve a considerable number of patients and duration of follow-up), the statistical thresholds guiding this decision should be prespecified and based on the need to maintain the desired overall type 1 error rate (typically 5%).[81,82] However, just because these guidelines are met does not mean a trial is automatically stopped,[83] particularly if the number of patients experiencing the endpoint is few. Additionally, a DSMB may recommend that a study continue even if the stopping boundary is crossed, if questions regarding the effects of treatment on important subgroups or secondary endpoints involving either efficacy or toxicity still remain.[83] A DSMB's recommendation to the investigators and sponsor to stop a trial for patient safety or efficacy is a difficult advisory decision that relies on the independence and impartiality of a DSMB. Some large veterinary RCTs have recently utilized DSMBs, interim analysis, or both including a trial that was prematurely halted due to safety concerns,[d] a trial in which interim analysis deemed it best to continue to its scheduled (and ultimately positive) endpoint,[24] and a trial that was prematurely halted for benefit after meeting stringent prespecified criteria.[4]

Finally, in addition to sound study design and analysis, independent reviewers of scientific journals are important arbiters of quality. Peer review should inject a healthy dose of skepticism to round out the perceptual bias of investigators toward their own results.[82] To be fair, there is always an asymmetry in how closely any potentially negative or discrepant result is scrutinized. The mere suggestion of ineffectiveness or harm elicits intense scrutiny, whereas similar levels of evidence pointing toward benefit often are ignored.[34] In general, reviewers should ensure that the benefits and risks of an intervention are clearly discussed, supported by the presented data, and are primarily derived from prespecified analysis and criteria.

## Conclusion

The age of prospective RCTs in veterinary medicine is fully upon us. There is an acute need to ensure the valid design and accurate reporting of RCTs and to maximally leverage the limited resources we have in the veterinary profession. Close attention to study endpoints, different forms of risk analysis, and issues of planning, monitoring, and reporting are needed. Many

insights can be drawn from the wealth of experience with RCTs in the human medical field and greater cooperation between biostatisticians in veterinary and human medicine and those performing the trials could achieve useful results. Simultaneously, efforts should be made to increase the training of veterinary students, generalists, and specialists in these areas. The issues covered here are just a small sample of the important considerations facing this new era of veterinary RCTs. Increased awareness and attention to these issues will move us faster and farther toward improved care of our patients.

## Footnotes

[a] https://www.clinicaltrials.gov/; accessed Feb 27, 2017.
[b] https://ebusiness.avma.org/aahsd/study_search.aspx; accessed Feb 27, 2017.
[c] http://vetalltrials.org; accessed Mar 16, 2017.
[d] Keene BW, Fox PR, Hamlin RL, Beddies GF, Keene TJ, Settje T, Treml LS. Efficacy of BAY 41-9202 (Bisoprolol oral solution) for the treatment of chronic valvular heart disease (CVHD) in dogs [Abstract]. Presented at the American College of Veterinary Internal Medicine Forum 2012, June 1, 2012, New Orleans LA.

## Acknowledgments

No grant support.

*Conflict of Interest Declaration*: Authors declare no conflict of interest.

*Off-label Antimicrobial Declaration*: Authors declare no off-label use of antimicrobials.

## References

1. Olby NJ, Muguet-Chanoit AC, Lim JH, et al. A placebo-controlled, prospective, randomized clinical trial of polyethylene glycol and methylprednisolone sodium succinate in dogs with intervertebral disk herniation. J Vet Intern Med 2016;30:206–214.

2. Schmitz S, Glanemann B, Garden OA, et al. A prospective, randomized, blinded, placebo-controlled pilot study on the effect of Enterococcus faecium on clinical activity and intestinal gene expression in canine food-responsive chronic enteropathy. J Vet Intern Med 2015;29:533–543.

3. Allstadt SD, Rodriguez CO Jr, Boostrom B, et al. Randomized phase III trial of piroxicam in combination with mitoxantrone or carboplatin for first-line treatment of urogenital tract transitional cell carcinoma in dogs. J Vet Intern Med 2015;29:261–267.

4. Boswood A, Haggstom J, Gordon SG, et al. Effect of pimobendan in dogs with preclinical myxomatous mitral valve disease and cardiomegaly: The EPIC study-a randomized clinical trial. J Vet Intern Med 2016;30:1765–1779.

5. Kristiansen VM, Pena L, Diez Cordova L, et al. Effect of ovariohysterectomy at the time of tumor removal in dogs with mammary carcinomas: A randomized controlled trial. J Vet Intern Med 2016;30:230–241.

6. Vollmar AC, Fox PR. Long-term outcome of Irish wolfhound dogs with preclinical cardiomyopathy, atrial fibrillation, or both treated with pimobendan, benazepril hydrochloride, or methyldigoxin monotherapy. J Vet Intern Med 2016;30:553–559.

7. Rausch-Derra L, Huebner M, Wofford J, et al. A prospective, randomized, masked, placebo-controlled multisite clinical study of grapiprant, an EP4 prostaglandin receptor antagonist (PRA), in dogs with osteoarthritis. J Vet Intern Med 2016;30:756–763.

8. Olby NJ, Vaden SL, Williams K, et al. Effect of cranberry extract on the frequency of bacteriuria in dogs with acute thoracolumbar disk herniation: A randomized controlled clinical trial. J Vet Intern Med 2016;31:60–68.

9. Zollers B, Wofford JA, Heinen E, et al. A prospective, randomized, masked, placebo-controlled clinical study of capromorelin in dogs with reduced appetite. J Vet Intern Med 2016;30:1851–1857.

10. Huhtinen M, Derre G, Renoldi HJ, et al. Randomized placebo-controlled clinical trial of a chewable formulation of amlodipine for the treatment of hypertension in client-owned cats. J Vet Intern Med 2015;29:786–793.

11. Sent U, Gossl R, Elliott J, et al. Comparison of efficacy of long-term oral treatment with telmisartan and benazepril in cats with chronic kidney disease. J Vet Intern Med 2015;29:1479–1487.

12. Blass KA, Schober KE, Li X, et al. Acute effects of ivabradine on dynamic obstruction of the left ventricular outflow tract in cats with preclinical hypertrophic cardiomyopathy. J Vet Intern Med 2014;28:838–846.

13. Geddes RF, Biourge V, Chang Y, et al. The effect of moderate dietary protein and phosphate restriction on calcium-phosphate homeostasis in healthy older cats. J Vet Intern Med 2016;30:1690–1702.

14. Short DM, Moore DA, Sischo WM. A randomized clinical trial evaluating the effects of oligosaccharides on transfer of passive immunity in neonatal dairy calves. J Vet Intern Med 2016;13949.

15. Pipkin KM, Hagey JV, Rayburn MC, et al. A randomized clinical trial evaluating metabolism of colostral and plasma derived immunoglobulin G in Jersey bull calves. J Vet Intern Med 2015;29:961–966.

16. Cohen ND, Slovis NM, Giguere S, et al. Gallium maltolate as an alternative to macrolides for treatment of presumed Rhodococcus equi pneumonia in foals. J Vet Intern Med 2015;29:932–939.

17. Schoster A, Staempfli HR, Abrahams M, et al. Effect of a probiotic on prevention of diarrhea and Clostridium difficile and Clostridium perfringens shedding in foals. J Vet Intern Med 2015;29:925–931.

18. Heller J. Epidemiological and statistical considerations for interpreting and communicating oncology clinical trials. Vet J 2015;205:233–237.

19. Giuffrida MA. Type II error and statistical power in reports of small animal clinical trials. J Am Vet Med Assoc 2014;244:1075–1080.

20. Di Girolamo N, Meursinge Reynders R. Deficiencies of effectiveness of intervention studies in veterinary medicine: A cross-sectional survey of ten leading veterinary and medical journals. PeerJ 2016;4:1649.

21. Giuffrida MA. Defining the primary research question in veterinary clinical studies. J Am Vet Med Assoc 2016;249:547–551.

22. Cordoba G, Schwartz L, Woloshin S, et al. Definition, reporting, and interpretation of composite outcomes in clinical trials: Systematic review. BMJ 2010;341:c3920.

23. Haggstrom J, Boswood A, O'Grady M, et al. Effect of pimobendan or benazepril hydrochloride on survival times in dogs with congestive heart failure caused by naturally occurring myxomatous mitral valve disease: The QUEST study. J Vet Intern Med 2008;22:1124–1135.

24. Summerfield NJ, Boswood A, O'Grady MR, et al. Efficacy of pimobendan in the prevention of congestive heart failure or sudden death in Doberman Pinschers with preclinical dilated cardiomyopathy (the PROTECT Study). J Vet Intern Med 2012;26:1337–1349.

25. Bernay F, Bland JM, Haggstrom J, et al. Efficacy of spironolactone on survival in dogs with naturally occurring mitral regurgitation caused by myxomatous mitral valve disease. J Vet Intern Med 2010;24:331–341.

26. Hogan DF, Fox PR, Jacob K, et al. Secondary prevention of cardiogenic arterial thromboembolism in the cat: The double-blind, randomized, positive-controlled feline arterial thromboembolism; clopidogrel vs. aspirin trial (FAT CAT). J Vet Cardiol 2015;17(Suppl 1):S306–317.

27. Group IEEW. Statistical principles for clinical trials: ICH harmonized tripartite guideline. Stat Med 1999;18:37.

28. Tomlinson G, Detsky AS. Composite end points in randomized trials: There is no free lunch. JAMA 2010;303:267–268.

29. Anker SD, Schroeder S, Atar D, et al. Traditional and new composite endpoints in heart failure clinical trials: Facilitating comprehensive efficacy assessments and improving trial efficiency. Eur J Heart Fail 2016;18:482–489.

30. Gomez G, Gomez-Mateu M, Dafni U. Informed choice of composite end points in cardiovascular trials. Circ Cardiovasc Qual Outcomes 2014;7:170–178.

31. Ferreira-Gonzalez I, Permanyer-Miralda G, Busse JW, et al. Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. J Clin Epidemiol 2007;60:651–657.

32. Royston P, Barthel FM, Parmar MK, et al. Designs for clinical trials with time-to-event outcomes based on stopping guidelines for lack of benefit. Trials 2011;12:81.

33. Freemantle N, Calvert M, Wood J, et al. Composite outcomes in randomized trials: Greater precision but with greater uncertainty? JAMA 2003;289:2554–2559.

34. Pocock SJ, McMurray JJ, Collier TJ. Statistical controversies in reporting of clinical trials: Part 2 of a 4-part series on statistics for clinical trials. J Am Coll Cardiol 2015;66:2648–2662.

35. Keech A, Simes RJ, Barter P, et al. Effects of long-term fenofibrate therapy on cardiovascular events in 9795 people with type 2 diabetes mellitus (the FIELD study): Randomised controlled trial. Lancet 2005;366:1849–1861.

36. Gerstein HC, Yusuf S, Bosch J, et al. Effect of rosiglitazone on the frequency of diabetes in patients with impaired glucose tolerance or impaired fasting glucose: A randomised controlled trial. Lancet 2006;368:1096–1105.

37. Ferreira-Gonzalez I, Busse JW, Heels-Ansdell D, et al. Problems with use of composite end points in cardiovascular trials: Systematic review of randomised controlled trials. BMJ 2007;334:786–793.

38. Brown PM, Anstrom KJ, Felker GM, et al. Composite end points in acute heart failure research: Data simulations illustrate the limitations. Can J Cardiol 2016;32:1356.e1321–1356.e1328.

39. Tong BC, Huber JC, Ascheim DD, et al. Weighting composite endpoints in clinical trials: Essential evidence for the heart team. Ann Thorac Surg 2012;94:1908–1913.

40. Lachin JM. Statistical considerations in the intent-to-treat principle. Control Clin Trials 2000;21:167–189.

41. International Conference on the Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. ICH Harmonised Tripartite Guideline: Statistical Principles for Clinical Trials E9 1998.

42. Gravel J, Opatrny L, Shapiro S. The intention-to-treat approach in randomized controlled trials: Are authors saying what they do and doing what they say? Clin Trials 2007;4:350–356.

43. National Research Council (US) Panel of Handling Missing Data in Clinical Trials. The Prevention and Treatment of Missing Data in Clinical Trials. The Prevention and Treatment of Missing Data in Clinical Trials. Washington, DC: National Academies Press; 2010.

44. Porta N, Bonet C, Cobo E. Discordance between reported intention-to-treat and per protocol analyses. J Clin Epidemiol 2007;60:663–669.

45. Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. N Engl J Med 1980;303:1038–1041.

46. Beckett RD, Loeser KC, Bowman KR, et al. Intention-to-treat and transparency of related practices in randomized, controlled trials of anti-infectives. BMC Med Res Methodol 2016;16:106.

47. Kruse RL, Alper BS, Reust C, et al. Intention-to-treat analysis: Who is in? Who is out? J Fam Pract 2002;51:969–971.

48. Sargeant JM, Thompson A, Valcour J, et al. Quality of reporting of clinical trials of dogs and cats and associations with treatment effects. J Vet Intern Med 2010;24:44–50.

49. Sargeant JM, O'Connor AM, Renter DG, et al. Reporting of methodological features in observational studies of pre-harvest food safety. Prev Vet Med 2011;98:88–98.

50. Schulz KF, Altman DG, Moher D, et al. CONSORT 2010 statement: Updated guidelines for reporting parallel group randomized trials. Ann Intern Med 2010;152:726–732.

51. Sargeant JM, O'Connor AM, Gardner IA, et al. The REFLECT statement: Reporting guidelines for randomized controlled trials in livestock and food safety: Explanation and elaboration. J Food Prot 2010;73:579–603.

52. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients: The need for risk stratification. JAMA 2007;298:1209–1212.

53. Kent DM, Alsheikh-Ali A, Hayward RA. Competing risk and heterogeneity of treatment effect in clinical trials. Trials 2008;9:30.

54. Ioannidis JP, Lau J. The impact of high-risk patients on the results of clinical trials. J Clin Epidemiol 1997;50:1089–1098.

55. Bewick V, Cheek L, Ball J. Statistics review 12: Survival analysis. Crit Care 2004;8:389–394.

56. Case LD, Kimmick G, Paskett ED, et al. Interpreting measures of treatment effect in cancer clinical trials. Oncologist 2002;7:181–187.

57. Bewick V, Cheek L, Ball J. Statistics review 11: Assessing risk. Crit Care 2004;8:287–291.

58. Varadhan R, Weiss CO, Segal JB, et al. Evaluating health outcomes in the presence of competing risks: A review of statistical methods and clinical applications. Med Care 2010;48:S96–105.

59. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989;81:1879–1886.

60. D'Agostino RB Sr, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: The Framingham Heart Study. Circulation 2008;117:743–753.

61. Campos M, Ducatelle R, Rutteman G, et al. Clinical, pathologic, and immunohistochemical prognostic factors in dogs with thyroid carcinoma. J Vet Intern Med 2014;28:1805–1813.

62. de Araujo MR, Campos LC, Ferreira E, et al. Quantitation of the regional lymph node metastatic burden and prognosis in malignant mammary tumors of dogs. J Vet Intern Med 2015;29:1360–1367.

63. van Rijn SJ, Hanson JM, Zierikzee D, et al. The prognostic value of perioperative profiles of ACTH and cortisol for recurrence after transsphenoidal hypophysectomy in dogs with corticotroph adenomas. J Vet Intern Med 2015;29:869–876.

64. Borgarelli M, Crosara S, Lamb K, et al. Survival characteristics and prognostic variables of dogs with preclinical chronic

degenerative mitral valve disease attributable to myxomatous degeneration. J Vet Intern Med 2012;26:69–75.

65. Hezzell MJ, Boswood A, Chang YM, et al. The combined prognostic potential of serum high-sensitivity cardiac troponin I and N-terminal pro-B-type natriuretic peptide concentrations in dogs with degenerative mitral valve disease. J Vet Intern Med 2012;26:302–311.

66. Chiavaccini L, Hassel DM. Clinical features and prognostic variables in 109 horses with esophageal obstruction (1992-2009). J Vet Intern Med 2010;24:1147–1152.

67. McConachie E, Giguere S, Barton MH. Scoring system for multiple organ dysfunction in adult horses with acute surgical gastrointestinal disease. J Vet Intern Med 2016;1276–1283.

68. Buczinski S, Boulay G, Francoz D. Preoperative and postoperative L-lactatemia assessment for the prognosis of right abomasal disorders in dairy cattle. J Vet Intern Med 2015;29: 375–380.

69. Kent DM, Rothwell PM, Ioannidis JP, et al. Assessing and reporting heterogeneity in treatment effects in clinical trials: A proposal. Trials 2010;11:85.

70. Koller MT, Raatz H, Steyerberg EW, et al. Competing risks and the clinical community: Irrelevance or ignorance? Stat Med 2012;31:1089–1097.

71. Haynes RB. Forming research questions. J Clin Epidemiol 2006;59:881–886.

72. DeAngelis CD, Drazen JM, Frizelle FA, et al. Clinical trial registration: A statement from the International Committee of Medical Journal Editors. JAMA 2004;292:1363–1364.

73. International Committee of Medical Journal Editors. Philadelphia PA: International Committee of Medical Journal Editors; 2016.

74. Broman K, Cetinkaya-Rundel M, Nussbaum A, et al. Recommendations to Funding Agencies for Supporting Reproducible Research. American Statistical Association; 2017.

75. U.S. Food and Drug Administration. Rare Diseases: Common Issues in Drug Development: Guidance for Industry. Rockville, MD: Center for Drug Evaluation and Research, U.S. Food and Drug Administration; 2015.

76. Pariser A. Rare Disease and Clinical Trials. Rockville, MD: Center for Drug Evaluation and Research, U.S. Food and Drug Administration; 2014.

77. Cornu C, Kassai B, Fisch R, et al. Experimental designs for small randomised clinical trials: An algorithm for choice. Orphanet J Rare Dis 2013;8:48.

78. Wittes J. Forming your phase III trial's data and safety monitoring board: A perspective on safety. J Investig Med 2004;52:453–458.

79. Ellenberg SS, Flemming TR, DeMets DL. Data Monitoring Committees in Clinical Trials: A Practical Approach. Hoboken, NJ: Wiley; 2002.

80. U.S. Food and Drug Administration. Guidance for Clinical Trial Sponsors: Establishment and Operation of Clinical Trial Data Monitoring Committees. Rockville, MD: U.S. Food and Drug Administration; 2006.

81. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35:549–556.

82. Hayward RA, Kent DM, Vijan S, et al. Reporting clinical trial results to inform providers, payers, and consumers. Health Aff (Millwood) 2005;24:1571–1581.

83. Pocock SJ, Clayton TC, Stone GW. Challenging issues in clinical trial design: Part 4 of a 4-part series on statistics for clinical trials. J Am Coll Cardiol 2015;66:2886–2898.