


REVIEW

What is the value of statistical testing of observational data?

Nick D. Jeffery BVSc, MSc, PhD, CertSAO, DipECVS, DipECVN, DSAS (Soft tissue), FRCVS¹ | Christine M. Budke DVM, PhD² | Guillaume P. Chanoit DEDV, MSc, PhD, DipECVS, DipACVS, FHEA, FRCVS³ 

¹Department of Small Animal Clinical Sciences, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, Texas, USA

²Department of Veterinary Integrative Biosciences, College of Veterinary Medicine & Biomedical Sciences, Texas A&M University, College Station, Texas, USA

³Small Animal Referral Hospital Langford Vets, University of Bristol, Bristol, UK

Correspondence

Guillaume P. Chanoit, Small Animal Referral Hospital Langford Vets, University of Bristol, Bristol, UK.
Email: g.chanoit@bristol.ac.uk

Funding information

No grant or other financial support was received.

Abstract

Statistical analysis of medical data aims to reveal patterns that can aid in decision making for future cases and, hopefully, improve patient outcomes. Large and bias-free datasets, such as those produced in formal randomized clinical trials, are necessary to make such analyses as reliable as possible. For a host of reasons, randomized trials are, unfortunately, relatively uncommon in veterinary medicine and surgery, implying that less ideal datasets (mostly observational data) must form the basis for much of our decision making regarding treatment of individual patients under our care. In this review, we first describe the common shortcomings of many observational veterinary datasets when viewed in comparison with their optimal counterparts and highlight how the deficiencies can lead to unreliable conclusions. We illustrate how many of the interpretative problems associated with observational data, predominantly various forms of bias, are not solved, and may even be exacerbated, by statistical analysis. We emphasize the need to examine summary data and its derivation in detail without being lured into relying upon *P* values to draw conclusions and advocate for completely omitting statistical analysis of many observational datasets. Finally, we present some suggestions for alternative statistical methods, such as propensity scoring and Bayesian methods, which might help reduce the risk of drawing unwarranted, and overconfident, conclusions from imperfect data.

1 | INTRODUCTION

Key aims of veterinary clinical research are to determine whether medical and surgical interventions are effective, to communicate this information to other veterinarians, and thereby provide evidence-based advice to owners. Ideally, evidence regarding comparative efficacy is derived from carefully designed clinical trials, in which

cohorts of patients are recruited prospectively, randomly allocated to different treatments, and outcomes are recorded by blinded observers. Such formal trials also include prestudy sample-size calculations to ensure sufficiently high power for the results to be robust (see <http://www.consort-statement.org/>).¹

Unfortunately, there are many reasons why it is not, and, for some conditions, never will be, possible to

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Veterinary Surgery* published by Wiley Periodicals LLC on behalf of American College of Veterinary Surgeons.

conduct randomized clinical trials (RCTs) for every veterinary procedure, including lack of clinical trial units (or even insufficient personnel to be able to ensure blinding), the extended timeframe needed to accrue the necessary number of cases, and widespread veterinarian resistance to randomization. Because of these obstacles, much veterinary clinical research relies instead upon analysis of observational, predominantly retrospective, data that do not meet the key design criteria for a reliable clinical trial (ie, random allocation of the intervention, inclusion of a comparison group, blinded recording of outcome). These studies can be helpful in clinical decision making but they unfortunately include numerous implicit pitfalls – notably a high risk of bias – and therefore a high risk of reaching erroneous conclusions. Such errors not only impede clinical veterinary progress but can also result in unnecessary morbidity and mortality, as has happened in human medicine in the past.² To minimize this hazard it is necessary to be aware of these pitfalls and the risk that they may be compounded by inappropriate statistical analysis.

As reviewers for *Veterinary Surgery* and other journals, we frequently encounter manuscripts containing many statistical tests, often apparently aiming to convince readers that one treatment is superior, or that one group of patients has a different prognosis, from another. Of course, these are important clinical questions but, as we discuss here, poorly designed statistical testing on observational datasets will not achieve that aim.

2 | BIAS AND CONFOUNDING IN CLINICAL OBSERVATIONAL DATA

Randomized clinical trials are designed to eliminate bias as far as possible. In this context, bias means systematic deviation of results or inferences from the truth. Methods to reduce bias include randomized and blinded allocation to treatment and assessment of outcome by a blinded observer (the term “blinded” metaphorically refers to the practice of drawing a blind over a window). By definition, observational studies do not include these features and are therefore more at risk of bias. Nevertheless, observational studies *can* provide valuable information, but great care is required in interpretation.

2.1 | Sources of bias in observational studies

A disadvantage of many veterinary observational studies is the use of retrospective data. Retrospective data are readily available and inexpensive to obtain but they

inevitably contain many more sources of bias. Bias can be classified in many different ways but here we list the categories used by Cochrane (www.cochrane.org): *selection bias*, *performance bias*, *detection bias*, *attrition bias*, *reporting bias*, and *other bias* (Table 1).

TABLE 1 Types of bias most encountered in clinical studies/trials. Adapted from: Chapter 8 Assessing risk of bias in included studies, Cochrane Handbook (https://handbook-5-1.cochrane.org/chapter_8/8_4_introduction_to_sources_of_bias_in_clinical_trials.htm)

Bias type	Definition	Example
Selection ^a	Systematic differences in baseline characteristics between groups	Nonrandom allocation Clinician Owner selection Selection by demographic
Performance ^a	Systematic differences between groups in overall medical care	Diagnostic investigation Intensity of follow up Prescribed medications
Detection ^a	Systematic difference between groups in how outcomes are determined	Nonblinded assessment (“observer bias”) Less complete examination of one group Different chronology between groups Recall bias (esp. case-control studies)
Attrition ^b	Systematic differences between groups in loss to follow up	High death rate in one group Incomplete capture of recovery in more slowly successful therapies
Reporting	Systematic differences between what is reported and what remains unreported	Incomplete disclosure of results

^aRandomized allocation, allocation sequence concealment and blinded outcome assessment (preferably also including blinding owners) will eliminate many of these biases in studies with large sample sizes.

^bMissing data are common in retrospective studies and constitute attrition bias for which it is difficult to adjust, because it is difficult to quantify, and it is often of unknown magnitude. Missing data often derive from cases that have unusual outcomes, such as individuals taking a long time to respond or having complications, and so omission is especially likely to cause bias.

2.2 | Confounding

Confounding factors are often associated with observational study groups. Unlike other types of bias that can result in the appearance of associations that are not true, confounding describes an association that is real but misleading (Figure 1). In the absence of randomization, it is possible, or even likely, that some aspects of baseline data will differ systematically between comparator groups. Confounding is said to occur if a variable is associated with both the exposure (ie, treatment) and the outcome, but is not on the causal pathway (from exposure to outcome). This confounding can happen because of direct associations between outcome, treatment effect, and confounder (Figure 1) but also, in an observational study, through an imbalance in the frequency of the confounder between the compared groups because of the way treatments are prescribed. This form of confounding is sometimes referred to as “confounding by indication.” For instance, following arytenoid lateralization surgery, a proportion of dogs in a specific hospital are admitted to the intensive care unit (ICU) and the remainder are kept in regular clinic wards. An observational study determines that there is a higher death rate in the dogs admitted to ICU. Again, admittance to ICU is a possible cause of death but it is much more probable that there is confounding by indication: dogs are not admitted to ICU randomly, and so are more likely to become ICU patients if they show various risk factors for poor outcome. That those dogs then have a higher proportion of deaths than those in regular wards does not indicate that ICU is unsafe! It is also important to note that such

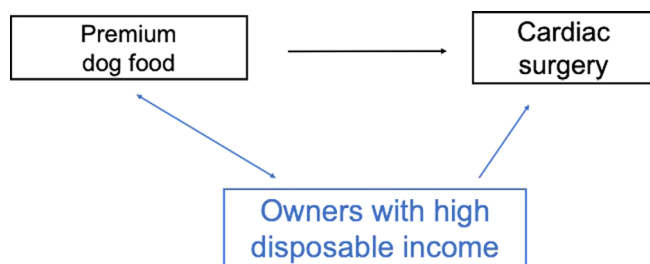


FIGURE 1 Example of confounding bias. In an observational study, it is found that a larger proportion of dogs undergoing valve replacement therapy are fed on premium dogfood than in a control group. Although it is possible, the simple explanation of causality (black arrow) is much less likely than the alternative explanation of confounding by owner disposable income (blue arrows). Owners of dogs presented for this type of expensive surgery will disproportionately have high disposable income and are therefore also more likely to feed their dogs on expensive dogfood. The risk of confounding in this example is clear but in many real-life examples the possibility is often much more easily overlooked

undermining of the assumption of random (and therefore equal) distribution of measured and unmeasured preintervention variables will lead to bias in outcomes of statistical testing (see below).

3 | STATISTICAL POWER, EFFECT SIZE AND SAMPLE SIZE

Power in statistics commonly refers to the probability of detecting an effect if it were truly there. However, it is less well recognized that low-power studies are also more likely to falsely “detect” effects that are not there.³ It is therefore probably more helpful to think of low power as producing results that are highly susceptible to the play of chance. For example, it is intuitive that small samples are much more likely to produce extreme results, such as a series of 100% heads in 4 rolls of the dice than in 40 rolls (or samples). Increased sample size is the most straightforward means to increase the power of a study.

Small data sets imply low statistical power (and therefore less reliable conclusions), so clinical research studies should be carefully planned beforehand to ensure that sample populations of appropriate size are acquired. Pre-study sample-size calculation need not be limited to experimental, prospective study designs but can also be included in retrospective studies. For instance, it would be possible to ask a question (before examining the data) about differences in serum alanine aminotransferase (ALT) activity between dogs with different types of liver shunt and analyze already available data. Data on standard deviations of this measurement are readily available and the researcher can specify the difference magnitude (in mean value) between groups it might be important to detect. Although these 2 statistics are sufficient for a sample size calculation it is also necessary to be cognizant of the clinical meaning of any detected difference. Although it is straightforward to calculate the sample size required to detect a difference of a specific magnitude between groups, it is also important to consider the clinical implications of finding such a difference. If the sought-for difference is so small as to be clinically unimportant then the sample size will be unnecessarily large. On the other hand, sample sizes calculated to detect very large differences may not be that helpful either because, first, very large differences are often implausible (it is medicine, not magic!) and, second, powering to detect only large differences implies the likelihood of overlooking smaller differences between treatment groups that may also be clinically meaningful. In this second instance it would be preferable to increase the sample size to ensure detection of any difference that might have clinical impact.

In RCTs, the power of the study is predetermined by the investigators and is usually set to be at least 80%-90%, meaning that, if there was truly to be an effect of the intervention (of the specified magnitude), there would, on average, only be a 10%-20% chance of overlooking it. It is notable that human RCTs with this level of power commonly enroll thousands of patients (depending on the effect size that is sought), hinting that the power of many veterinary observational studies, although largely unknown, is probably extremely low. As a comparison, post hoc estimates of power of “blue sky” bench research in neuroscience typically range from 8% to 30%,³ and it is likely that many observational studies without sample size calculations in veterinary medicine would have similar power, meaning that the chance of reliably identifying true effects is low. This lack of power is a likely explanation for the many contradictory conclusions drawn from published studies on risk factors for a poor outcome after a specific surgery even when findings have been designated “statistically significant.” The factors that are “detected” in such low-power studies more likely simply reflect the play of chance in that specific dataset and may be strongly influenced by selection and allocation bias associated with lack of randomization.

4 | IMPORTANCE OF EFFECT SIZE

In hypothesis testing (see below), effect size, power, and sample size are interrelated. If an effect size is large and is detected with little variability within the study population, then the power of a study may be high even with a small sample size. Although choice of study power is a key determinant of necessary sample size, it is also dependent on the magnitude of the intervention effect and its variability. In clinical research, the effect size is the magnitude of treatment effect on the outcome of interest, usually presented as a summary value, such as mean, median, odds ratio, etc., together with an estimate of how precisely that summary figure is known. For instance, an effect size might be that the odds ratio for surgical wound infection for dogs undergoing standard open surgery versus laparoscopic ovariectomy is 2.2 (95% CI: 0.8-3.6). Alternatively, the serum ALT activity in dogs with intrahepatic portosystemic shunts might be a mean of 40 units higher than that in dogs with extrahepatic shunts, with a standard deviation of 20.

Although infrequently used in veterinary medicine, effect sizes in human medicine, especially clinical trials, are often summarized as the “number-needed-to-treat” (NNT) to achieve a specific end point. This measure is valued because it translates study results into a readily useable answer for everyday practice, although it is

necessary for it to be interpreted in conjunction with the baseline risk of the event of interest. A recent study on treatment of intervertebral disc herniation in dogs provides a veterinary example: Martin et al. (2020)⁴ calculated that 14 dogs (ie, the NNT) presenting “deep pain-positive” needed to be taken to surgery on the same day (rather than the next day) for one to not become “deep pain-negative” within the following 24 hours. They also reported a wide 95% confidence interval for the NNT (of 7-106) indicative of considerable uncertainty about the magnitude of effect.

5 | NULL HYPOTHESIS STATISTICAL TESTING OF OBSERVATIONAL DATA

5.1 | What does null hypothesis testing detect?

Conventional null hypothesis statistical testing aims to evaluate how well the results fit with the null hypothesis of no difference between tested groups, and is formally stated as: “*If the null hypothesis is correct, how likely would it be to obtain these, or more extreme, results were the experiment to be repeated an infinite number of times?*” However, this statistical testing relies upon many assumptions, one of which is that, for a study evaluating a therapeutic intervention, the 2 sample groups have been randomly allocated to different treatments. In an RCT, assuming that the trial is conducted properly, differences in outcome can then reliably be attributed to differences in treatment.

In contrast, most manuscripts submitted to veterinary journals do not report findings from an RCT. It is therefore often questionable whether the implicit assumptions inherent in statistical testing have been met. For these reasons, care must be taken when interpreting statistical tests applied to observational data. The *P* value that results from conventional null hypothesis statistical testing is an indication of the probability that the data are compatible with the null hypothesis. In such testing, the null hypothesis assumes that there is no systematic difference between the 2 compared groups, *apart from the factor of interest* (e.g., surgical therapy). This assumption is clearly contravened in most sets of observational data because of the many sources of bias. For instance, the treatment groups might have been selected in some way (ie, selection bias – see above) that is inconsistent with the null hypothesis, or are distorted by inadvertent confounding with other important factors. This implies that there is frequently an a priori reason for differences between groups that has nothing to do with the

treatment under investigation. The extent to which an observational dataset is subject to bias (e.g., assessment of outcome by a nonblinded observer) is a matter of interpretation by the reader and cannot be unraveled by statistical testing.

It is frequently assumed in conventional statistical testing that a P value of less than a specific value (conventionally $P < .05$) can be used to distinguish real or important effects from false or meaningless effects. This is patently absurd, for many reasons, not least that the value of .05 is largely arbitrary. Fisher (an originator of the P value) simply suggested a variable reaching that level of “significance” was an interesting finding, not a delineator between true and false.^{5,6} There is always a need to look at the results in more detail, including the absolute size of the difference between groups, how precisely that difference has been estimated, and what that difference (or range of difference) would mean clinically. The P value is dependent not only on the size of the difference (and its variability) but also on the number of samples that have been analyzed. Large samples can therefore detect small differences that may be clinically unimportant.

5.2 | Using statistical testing with multiple variables

Statistical testing has most value when it has been planned into a study from the beginning, with carefully preplanned and limited analyses and appropriate sample sizes calculated beforehand. Possible exposure (or confounding) variables should also be considered during the study design process. For example, in a comparison of postoperative infection rates between 2 different surgical techniques that are not randomly allocated, it might also be important to consider the age and weight of each patient as well as other possible factors. However, when data are analyzed in this way there is “fragmentation” of the individuals into an often large number of possible categories, meaning that each category may contain few – or even zero – individuals. For instance, if we are examining the effects of treatment, but also allowing for effects of age and weight in 3 categories each, plus experienced versus resident surgeon, we now have 36 ($2 \times 3 \times 3 \times 2$) individual categories. This then implies that, unless the dataset is large, many of these categories will contain few data points, meaning that estimates of the effect of specific variables (and even more so, specific combinations of variables) will become highly susceptible to chance effects. In a low power study in which there have been few events (e.g., infection, death) recorded and many variables are entered into the

regression equation, it is highly possible that the small number of events will associate with one of the multitude of variables simply by chance. Indeed, this alone is likely to account for the large number of variables that have been associated with poor outcomes after (for instance) treatment of portosystemic shunt in dogs and the fact that there are few reports that agree in their conclusions. Similarly, there are many conflicting reports regarding whether splenectomy is a risk factor for gastric dilatation volvulus (GDV) and therefore uncertainty about whether to recommend routine gastropexy following splenectomy.⁷ For both study questions, almost all relevant studies are of low power and so the results of each of them is unreliable (although some may also, by chance, identify an important risk factor), thereby generating an (artificial) conflict that can be confusing and demoralizing for clinicians.

When entering a series of variables into a logistic regression analysis, it is important to keep focus on estimating the effect of the primary variable of interest and ensure that there are – by a rule-of-thumb – at least 10 outcome events (e.g., infections) for each entered variable (and even this number is not necessarily sufficient to ensure adequate power). To fulfill this objective, it is necessary to define, before the study is commenced, a clear clinical question and hypothesis. Addressing multiple hypotheses is problematic because it is a reflection that the study was not hypothesis driven but more exploratory driven, thereby implying that the conclusions should be treated with more caution. In that respect, the “patient/population, intervention, comparison and outcomes” (PICO) model is very useful to design the study question (https://libguides.mssm.edu/ebm/ebp_pico). Even in this scenario there may be better ways (e.g., the NNT described previously or the alternative analyses described below) to examine these data than traditional null hypothesis significance testing because the effect size is a more important result. The key thing that we want to know as clinicians is how effective a therapy is, not necessarily whether one treatment is “statistically superior” to another.

5.3 | Multiplicity

Using a P value of .05 implies that there will, on average, be 5 out of 100 (5%) “significant” findings purely due to chance, *even when the null hypothesis is correct*. Whilst it is straightforward to deal with the possibility that, at a P value of .05, 1 of 20 significant outcomes will result purely from chance, it does rapidly become more problematic if multiple comparisons are made, because it causes an accumulation of false positives. For instance, if

10 comparator tests are done and $P < .05$ is used as an alert for an interesting result, then there is a 50% chance of at least one P value being $<.05$ through chance alone, even when all the assumptions for the test have been met and *the null hypothesis is correct*. Multiple testing of this type also suggests that the researcher did not have a clear hypothesis in mind when conducting the analysis because otherwise there would not be the multiple hypotheses implied by the multiple statistical testing or worse, hypotheses have been derived from the results (ie, “hypothesizing after the results are known,” also known as “HARKing,” which is recognized as a form of research misconduct).⁸

Multiple testing for exploratory purposes is, of course, an important part of scientific discovery, but it is critical that it is presented as such and not as hypothesis-testing work, because otherwise inappropriate clinical decisions might be made. There are several methods for dealing with the risk of type I error associated with multiple testing, the most well known of which is the Bonferroni correction, which is a mathematical solution that can be applied to P values after testing. However, it does have many drawbacks and can be overly conservative. In addition, it is often confusing to read manuscripts in which the correction has been applied. An alternative method, the Benjamini-Hochberg procedure,⁹ is probably more useful for determining which in a series of exploratory statistical analyses might be worth pursuing further.

6 | RECOMMENDATIONS FOR ANALYSIS OF OBSERVATIONAL DATA

Thus far, we have made many complaints – but how can these problems be avoided? The first question might be whether statistical analysis adds anything useful to what might be gleaned simply by closely examining the data. Often, the statistical analysis adds nothing useful and might even lead readers astray. For instance, when confronted with a set of data *without* statistical analyses or P values, readers will usually examine in more detail how the data were collected, whether there were major differences between groups, and question more closely whether comparisons are valid or might be compromised by bias or confounding etc. Often, if we are interested in whether there is a major difference in observational data outcomes between 2 comparator groups then the P value will add little, in the absence of confirmatory data (which, by definition, would not be available). Similarly, if the assumption is made that the datasets were collected with limited bias (which statistical testing cannot confirm), readers would naturally recommend the treatment

with the best outcome, whether or not a P value is attached. As discussed above, the P value cannot be used to confirm that there is no bias, or limited bias, or confounding in the data.

Lastly, as an alternative there is Bayesian analysis, which evades the problems associated with repeated testing and, in contrast to null hypothesis testing, can be used to evaluate the strength of evidence *in favor* of a specific statistical model, including the null. Bayesian analysis relies on a very simple concept: that the likelihood of a specific hypothesis being true depends upon the amount of data in support. This type of conclusion is what many researchers want null hypothesis statistical testing to provide, even though it cannot.

7 | ASSESSING STATISTICAL ANALYSES FOR OBSERVATIONAL STUDIES IN PUBLISHED MANUSCRIPTS

The first recommendation for evaluating statistics associated with observational studies is to assess the results as a weighing up of evidence. It is important to consider that observational, especially retrospective, data will produce evidence that is “gray” but of variable shades. It will almost always be useful information but it would be rare for this type of study alone to definitively answer a research question. The process of evaluation is to determine whereabouts on the gray scale the evidence lies. Scrutiny of the article will enable a balanced conclusion to be drawn, bearing in mind the many errors that have been made in human medicine from overreliance on analysis of observational data.

The first part of this process is to examine the materials and methods section to determine how likely it is that fair comparisons between groups can be drawn. For studies looking at therapeutic interventions, this is best done by comparison of the methods with the ideal of the RCT (as described in the CONSORT [Consolidated Standards of Reporting Trials] guidelines- www.consort-statement.org) and examining the summaries of the demographics of the compared patient populations, which is usually contained in the results section.

The second recommendation is to concentrate on the data (usually in the form of summary statistics) rather than the inferential statistical tests. Researchers therefore need to make sure that such information is available in a table in the manuscript itself and, preferably, also include the raw data as supplementary material. In scrutinizing the results, the most important aspect is to examine the effect size and the precision of its estimation. How much difference is there between the tested groups? Is this

difference enough to be meaningful when dealing with real-life patients? For instance, an effect size of 5 beats per minute when comparing the effect of 2 drugs on heart rate is not likely to be important.

As an associated assessment, it is necessary to look at the precision of the estimate of effect. This can be examined by looking at the estimates of variation, often most usefully assessed by the 95% confidence intervals. Low power studies may suggest evidence of a treatment effect – say a point estimate of an odds ratio of 5.4 – and even have a 95% CI that does not incorporate the null value (1.0), suggesting statistical significance at $P < .05$. However, if the 95%CI values for the odds ratio is wide – say 1.1 to 10.9 – it would suggest that the study has low power and so the magnitude of effect is rather uncertain. On the other hand, the data are compatible with the possibility that the effect might be as large as an odds ratio of nearly 11, which would almost certainly be of clinical importance.

Lastly, it is preferable to not take too much notice of a $P < .05$ unless the study is appropriately powered, and with limited evidence of bias and confounding – for instance if assured of randomization of treatment. It is surprisingly easy to be led astray by statistical testing and jump to the easy conclusion. Frequently, consideration of the raw data frees your mind to look at the real story.¹⁰ For instance, if confronted with a $P < .05$ for comparison of death rates between medical and surgical therapy for a specific condition, it is often useful to probe the details of exactly how allocation of those 2 treatment options was determined. This choice is usually a clinician's recommendation, and so is likely biased. For these reasons, it is important to recognize that the value of a study might not be enhanced by statistical testing, and may even have detrimental effects in determining the true value of published data.

8 | ALTERNATIVES TO STANDARD ANALYTICAL TECHNIQUES

Although null hypothesis statistical testing has dominated scientific and medical literature for many decades, it does have a multitude of shortcomings, many of which are highlighted here. In medicine and surgery, it is usually important to have an idea of *how reliably we know* the comparable effectiveness of 2 competing therapies, and conventional statistical testing is not good at providing that information from observational data. Statistics, such as mean, standard deviation, median survival, and odds ratios, provide summary information on effect size but are often difficult to interpret on their own in observational studies because of the numerous possible confounding factors.

Although it must be emphasized that *statistical methods cannot extract bias* from data, there are some recent developments in statistics and in the availability of statistical software that might aid in producing more realistic and clinically meaningful estimations of effect in observational studies that move away from the rigid bright-line decision making that is often associated with traditional null hypothesis statistical testing. For instance, propensity scoring¹¹ and related statistical methods provide a means by which 2 observational groups can be compared, while allowing for numerous covariates (possible confounders and effect modifiers) to be included in the analysis. This type of analysis aims to reduce bias effects by including the various factors (covariates) that might determine allocation, outcome, or both. Notably, the results focus on effect sizes, rather than whether a treatment is “significantly different” from another. As an example, a comparison of the incidence of wound breakdown following suture or wire closure of sternotomy incisions in dogs was able to conclude that there was likely little difference in probability of failure between methods.¹² However, the 95% confidence intervals on the estimate indicated that the data were compatible with the possibility that there might be as much as a 10% lower rate of complications associated with sutures. This result is clearly important information for a clinician: either method of closure is an appropriate choice but if there is any uncertainty it is probably best to choose suture. Such a conclusion is of far more value than just saying that one therapy is “significantly” better than another – especially because, as noted above, significance is dependent upon the sample size. This “treatment effects” analysis provides an automatic limit to interpretation because of the inclusion of the 95% CI, which is dependent upon the sample size (larger sample corresponds with narrower 95%CI).

Similar emphasis on effect size and the likelihood of achieving such effects is also an inherent component of Bayesian analyses. The process involves examining the “posterior probability” (ie, how likely the hypothesis is to be true *after* the experiment), having started with a “prior probability” (ie, how likely the hypothesis was *before* the experiment). The ratio between these probabilities is determined by the weight of evidence provided by the experiment compared with what was available before. Until recently, the practical computational problems associated with these calculations and the lack of user-friendly software has hindered the widespread adoption of Bayesian analysis but, fortunately, these obstacles are beginning to be broken down. Appropriate software is becoming more widely available, such as that provided by the R Foundation (<https://www.r-project.org/foundation/>) (free of charge), Stata (StataCorp, College Station, TX; paid subscription required), or JASP (Jeffreys's Amazing Statistics

Program) (<https://jasp-stats.org/>). The latter might be especially attractive to veterinary surgeons wishing to dip their toes into Bayesian analysis because it is available as a free download and comes with an extensive instruction manual (<https://jasp-stats.org/2020/05/19/bayesian-inference-in-jasp-a-new-guide-for-students/>). This software can carry out straightforward Bayesian analyses that will often provide more directly useful statistical conclusions than standard null-hypothesis statistical testing.

A major advantage of Bayesian analysis is that it is not hindered by repeat testing because the problems with multiplicity mentioned above do not apply.¹³ It is therefore possible, and even desirable, to repeatedly test the data as it is collected. This method is commonly used in modern human clinical trials. As an outcome, it will produce a point estimate of effect size, together with a “credible interval” (the Bayesian equivalent of the confidence interval). As would be expected, with little data the credible interval will be wide, but then it narrows as more data are accrued. An example of how this can be valuable is provided by the analysis of effects of durotomy in dogs published in *Veterinary Surgery*¹⁴ in which the recovery to walk again after becoming deep pain negative following thoracolumbar intervertebral disc herniation and undergoing durotomy was estimated at 71% with a 95% credible interval of 52% to 87%. It was also possible to determine how likely this procedure was to produce an outcome superior to selected cutoff points, such as the ~55% recovery expected after routine decompressive surgery.¹⁵

The question of prior probability is simultaneously both more and less complicated! Bayesian analysis is undoubtedly more powerful if the prior probability can be computed accurately but, in real life this does remain difficult. For instance, what is the expected probability of a normal lifespan in a dog treated conservatively for portosystemic shunt, and at what level of certainty do we know this information? The problem for both these items is that the studies to provide this information are observational and may not reflect the true value and so it may be difficult to defend (in a publication) a specific choice of prior probability. The simple, and commonly applied, way out of this dilemma is to simply use what is referred to as a “noninformative prior.” This means applying a mathematical formula that implies that we just do not know what the outcome of the experiment is likely to be. This, of course, does weaken the impact of being able to apply previous knowledge, and implies that the outcome will more closely match that determined by conventional null hypothesis testing. However, it is more likely to find favor with readers and has the benefits of providing “Bayesian answers” (ie, how *likely* is the specific tested hypothesis to be true based on these data?) as contrasted with answers provided by null hypothesis

testing (ie, how *unlikely* is the null hypothesis?). As mentioned above, Bayesian analysis also permits further experimental data to be incorporated into the analysis as it is accrued, thereby generating progressively more precise estimates of effect (or lack of effect).

9 | CONCLUSION

It is important to be aware of the limitations and spurious results that can be generated by analysis of (especially, small) clinical observational datasets and the potential errors in patient care that they may induce. We need to be realistic and acknowledge that large RCTs are not always going to be possible for every condition but, instead of repeatedly falling into the same traps (as our colleagues in human medicine have previously done), we should consider redefining which statistical tests and methodology are most useful in guiding clinical decisions from observational data sets.

It is also important to realize that some of the choices in methodology of veterinary clinical research (eg, choosing observational data analysis, single center analysis, or short case series) might be driven by nonscientific imperatives such as satisfying board credentials that limit the appetite of clinicians for longer term prospective studies. It may be timely to reconsider how these imperatives influence the type and impact of clinical veterinary research and whether adjustments could then lead to greater opportunities to conduct larger, multicenter, observational retrospective or prospective, especially RCT, investigations.

ACKNOWLEDGMENTS

Author Contributions: Jeffery ND, BVSc, PhD, MSc, CertSAO, DipECVS, DipECVN, DSAS (soft tissue), FRCVS: Manuscript writing, original idea, and study design; Budke CM, DVM, PhD: Manuscript writing, expert in veterinary epidemiology; Chanoit GP, DEDV, MSc, PhD, DipECVS, DipACVS, FHEA, FRCVS: Manuscript writing, original idea, and study design.

CONFLICT OF INTEREST

The authors declare no conflicts of interest related to this report.

ORCID

Guillaume P. Chanoit  <https://orcid.org/0000-0002-7414-6403>

REFERENCES

1. Wittes J. Sample size calculations for randomized clinical trials. *Epidemiol Rev.* 2002;24:39-53.

2. Altman DG. The scandal of poor medical research. *BMJ*. 1994; 308:283-284.
3. Button KS, Ioannidis JP, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013;14:365-376.
4. Martin S, Liebel FX, Fadda A, Lazzarini K, Harcourt-Brown T. Same-day surgery may reduce the risk of losing pain perception in dogs with thoracolumbar disc extrusion. *J Small Anim Pract*. 2020;61:442-448.
5. Pearce SC. Introduction to Fisher (1925) statistical methods for research workers. In: Kotz S, Johnson NL, eds. *Breakthroughs in Statistics*. Springer Series in Statistics. Springer; 1992:59-65.
6. Wagenmakers E-J. A practical solution to the pervasive problems of p values. *Psychon Bull Rev*. 2007;14:779-804.
7. Harris O, Gordon-Evans WJ. Current evidence supporting simultaneous prophylactic gastropexy in canine patients undergoing complete splenectomy. *Vet Evid J*. 2021;6(4):1-10.
8. Andrade C. HARKing, cherry-picking, P-hacking, fishing expeditions, and data dredging and mining as questionable research practices. *J Clin Psychiatry*. 2021;82:20f13804.
9. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing. *J R Stat Soc B*. 1995;57:289-300.
10. Jeffery N. Liberating the (data) population from subjugation to the 5% (P-value). *J Small Anim Pract*. 2015;56:483-484.
11. Burton W, Drake C, Ogeer J, et al. Association between exposure to Ehrlichia spp. and risk of developing chronic kidney disease in dogs. *J Am Anim Hosp Assoc*. 2020;56:159-164.
12. Pilot M, Lutchman A, Hennes J, et al. Comparison of median sternotomy closure-related complication rates using orthopedic wire or suture in dogs: an observational treatment effects analysis. *Vet Surg*. 2022. doi:10.1111/vsu.13846. Online ahead of print.
13. Wagenmakers EJ, Marsman M, Jamil T, et al. Bayesian inference for psychology. Part I: theoretical advantages and practical ramifications. *Psychon Bull Rev*. 2018;25:35-57.
14. Jeffery ND, Mankin JM, Ito D, et al. Extended durotomy to treat severe spinal cord injury after acute thoracolumbar disc herniation in dogs. *Vet Surg*. 2020;49:884-893.
15. Langerhuus L, Miles J. Proportion recovery and times to ambulation for non-ambulatory dogs with thoracolumbar disc extrusions treated with hemilaminectomy or conservative treatment: a systematic review and meta-analysis of case-series studies. *Vet J*. 2017;220:7-16.

How to cite this article: Jeffery ND, Budke CM, Chanoit GP. What is the value of statistical testing of observational data? *Veterinary Surgery*. 2022; 51(7):1043-1051. doi:10.1111/vsu.13845