Published November 10, 2016

Interpreting statistics from published research to answer clinical and management questions[1]

B. J. White, R. L. Larson[2] and M. E. Theurer

Beef Cattle Institute, Kansas State University, Manhattan 66506

ABSTRACT: Appropriate statistical analysis is critical in interpreting results from published literature to answer clinical and management questions. Internal validity is an assessment of whether the study design and statistical analysis are appropriate for the hypotheses and study variables while controlling for bias and confounding. External validity is an assessment of the appropriateness of extrapolation of the study results to other populations. Knowledge about whether treatment or observation groups are truly different is unknown, but studies can be broadly categorized as exploratory or discovery, based on knowledge about previous research, biology, and study design, and this categorization affects interpretation. Confidence intervals, *P*-values, prediction intervals, credible intervals, and other decision aids are used singly or in combination to provide evidence for the likelihood of a given model but can be interpreted only if the study is internally valid. These decision aids do not test for bias, study design, or the appropriateness of applying study results to other populations dissimilar to the population tested. The biologic and economic importance of the magnitude of difference between treatment groups or observation groups as estimated by the study data and statistical interpretation is important to consider in clinical and management decisions. Statistical results should be interpreted in light of the specific question and production system addressed, the study design, and knowledge about pertinent aspects of biology to appropriately aid decisions.

Key words: confidence interval, inferential statistics, literature, *P*-value, research

 J. Anim. Sci. 2016.94:4959–4971
doi:10.2527/jas2016-0706

INTRODUCTION

Appropriate statistical analysis is critical for interpreting results from published literature to address a specific clinical or management question. Descriptive statistics are used to summarize study data; however, describing the data collected from a group of experimental or observational units in a study does not allow one to make conclusions beyond the study population. Inferential statistics provide a mechanism to use the study data to evaluate hypotheses and aid decision-making applied to a larger, unmeasured population.

In most studies, the outcome of interest differs numerically between treatment or observation groups, but the reader is most likely interested in determining if the observed numerical difference is an accurate description of the wider population or if the findings occurred by chance. Researchers recognize that one sampling of a few representative animals in a population is not expected to be the same as another sampling of the same population and that neither sampling is a perfect depiction of the entire population; hence, a numerical difference between treatment groups or observation groups has some probability of being due to chance. Inferential statistics uses estimates of the outcome of interest, estimates of precision (e.g., SE or confidence interval [**CI**]), and test statistics to aid the interpretation of study data in relation to the study hypotheses. Reading the scientific literature does not require an intricate understanding of how to perform appropriate statistics for every potential study design; however, a basic knowledge of how to assess the internal validity of studies and interpret statistical results is important for making inferences from published research to guide decision-making in animal and veterinary

[1]Funding for this project provided in part by Merial Limited, Duluth, GA 30096.

[2]Corresponding author: rlarson@vet.k-state.edu

Received June 7, 2016.
Accepted August 25, 2016.

sciences. The objective of this article is to describe the role of statistics in interpreting research and how to use statistical evidence to make inferences when answering clinical and management questions.

TYPES OF VARIABLES

Understanding the type and distribution of each study variable is important to determine the most appropriate statistical test. Variables fall into 2 main types, quantitative or qualitative, with multiple distinctions within those categories. Readers should investigate the type of variables being evaluated to assess if appropriate analyses were performed.

Quantitative continuous measurements can take on any numerical value. For example, BW is a continuous variable because animal weight can be reported in any increment allowed by the precision of the scale used (1 kg, 1 g, 1 mg, etc.). Quantitative discrete measurements such as litter size or counts of events must be whole numbers. Quantitative measurements can be made objectively (vs. subjectively) and the difference between each quantitative measurement increment is exactly the same (e.g., the weight difference between 1 and 2 kg is exactly the same as the difference in weight between 21 and 22 kg).

In contrast, descriptions are used for qualitative variables (dead vs. alive, sick vs. healthy, male vs. female, BCS, etc.). Sometimes scoring systems such as clinical illness score (**CIS**), lameness score, or BCS are reported as numerals, but, in fact, they are qualitative, not quantitative. Clinical illness scoring systems of feedlot cattle commonly include a score ranging from 0 to 4 with a score of 0 indicating apparently healthy animals and a score of 4 representing severely ill animals (Perino and Apley, 1998). These qualitative outcomes are not continuous or discrete numbers because each qualitative descriptive interval is not the same (e.g., the difference in disease severity between animals with CIS of 1 vs. 2 is not the same as the difference in disease severity between animals with CIS of 3 vs. 4). Scoring systems are commonly ordinal (ordered), as there is a relationship of severity or amount as the scale increases or decreases, but these scoring systems are still qualitative because they are being used to provide descriptions, not measurements. A method to determine whether scoring systems are qualitative or quantitative in nature is to determine if the scale could have been assigned letters instead of numbers and still have the same interpretation. If letters could be assigned to the scale, then the scoring system is qualitative.

Although reliability and agreement is important for all data collection methods, when the data is subjective, it is particularly important that the scoring system be validated to provide the reader an estimate of both the intra- and inter-rater agreement and reliability (Kottner et al., 2011). In addition, if ordinal data is collected and then levels within the scoring system are merged to create binary classifications for analysis, the justification for which levels are merged should be clearly stated and based on demonstrated improvements in reliability or agreement (Schlageter-Tello et al., 2014). Without a clear description and validation of the data collection method, the reader cannot interpret reported associations.

It is important to recognize that different statistical tests are used with qualitative and quantitative data. As an example, a research study could be conducted to compare the impact of treating cattle for bovine respiratory disease with 2 different antimicrobials (treatments A and B), and the outcomes of interest are ADG, mortality risk, and CIS. These outcomes represent different variable types because ADG is a continuous variable, mortality is binary data that can be only "yes" or "no" for an individual animal, and CIS is a qualitative variable that can assume only a limited number of ordinal descriptions. Figure 1 provides examples of each distribution for ADG, mortality risk, and CIS. Statistical software cannot detect if an entered numeral is a qualitative description rather than a quantitative measurement; therefore, statistical packages will run inappropriate calculations and report inaccurate results if the researcher does not recognize this common mistake. The same statistical methods should not be used to evaluate both CIS and ADG even if both are recorded using numerals. Additionally, it is improper to calculate a mean for qualitative data, such as CIS (for example, reporting a mean CIS of 1.3 for treatment A and 2.7 for treatment B). The goal of presenting a summary statistic is to represent the population as a whole. Reporting a qualitative variable such as CIS as having a mean of 1.3 or 2.7 is nonsensical because these values would not represent any actual animals, as study subjects could not receive a score between mutually exclusive categories. Medians should be presented as a summary statistic of central tendency for data where nonparametric tests will be used for inference testing.

Studies of almost any size result in numerical differences in the outcomes of interest. The function of inferential statistics is to quantify the likelihood that if the treatment or risk factor had no effect, a difference as great as or greater than that observed in the study would be due to chance and, by inference, some qualitative level of confidence that the difference observed was due to the treatment or risk factor of interest. Expanding on the preceding bovine respiratory disease treatment example, the ADG for cattle receiving treatment A could have been 1.5 kg/(animal·d), with a morbidity (abnormal CIS) risk of 21%, and a mortality risk of 0.6%; cattle receiving treatment B could have an ADG of 0.5 kg/

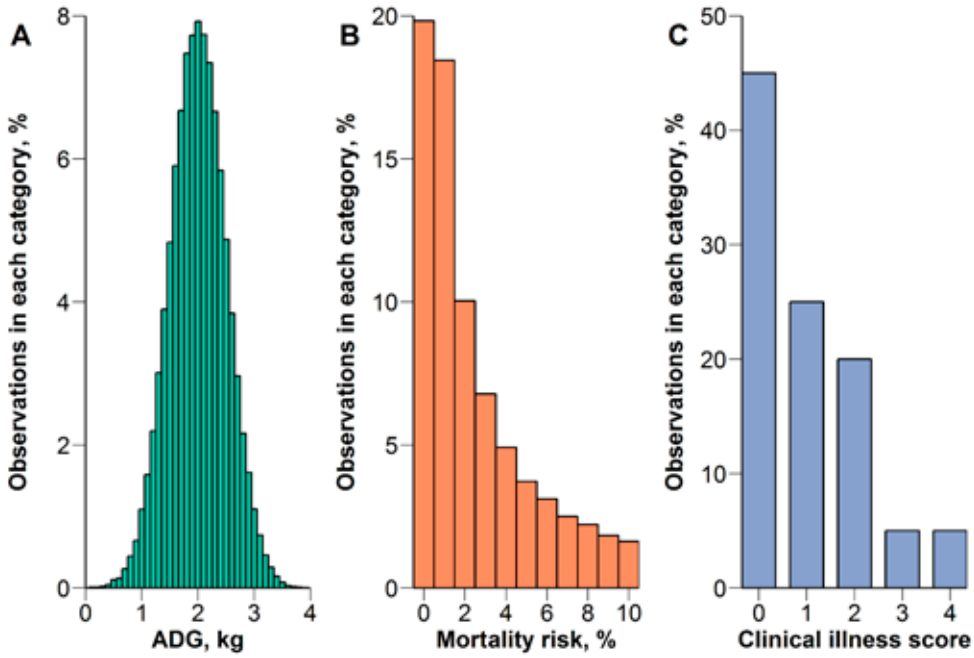

Figure 1. Examples of common outcome variables with normal (ADG; A), binomial or skewed (mortality risk; B), and discrete (clinical illness scores; C) data distributions.

(animal·d), a morbidity risk of 39%, and a mortality risk of 1.5%. The raw data supports the hypothesis that treatment A offers health and performance advantages, but it is impossible to determine whether the observed difference in the raw data is most likely associated with chance or a treatment effect. Although we can safely conclude that cattle receiving treatment A did, in fact, have better growth and health performance than cattle receiving treatment B in the sample population, our actual clinical question is, "Will cattle receiving treatment A have better growth and health performance than cattle receiving treatment B in future populations?" Or, stated another way, "Is the direction and magnitude of the numerical difference between treatment groups we observed in the experiment likely just due to chance?"

Traditional null hypothesis inference testing methodology will provide the probability (*P*-value) that if the model assumptions are true (including, most commonly, that the treatments were truly not different), differences as great as or greater than that observed in the study would occur by chance. Interpreting potential outcome variable differences is based on understanding the underlying variable characteristics. Weight gain in this study is monitored by a continuous variable with an expected normal (bell-shaped) distribution in the entire population (Fig. 1A). In other words, the expected distribution of animals' weight gain within the study population could be described by the mean ADG for the group as the center of the bell-shaped curve and the variation of individual animal ADG within the population as the width of the bell-shaped curve as described by the term standard deviation. A normal distribution has outcomes from animals in the population distributed evenly on both sides of the mean with approximately 66, 95, and 99% of the population being located within 1, 2, and 3 standard deviations of the mean, respectively. Therefore, the statistical test must evaluate the likelihood that mean observations as extreme as 0.5 and 1.5 kg/(animal·d) could be drawn from a single population. In contrast to ADG, mortality risk assumes a much different distribution and is often skewed to the right (Fig. 1B; Theurer et al., 2015). Using a scoring system such as CIS results in discrete, ordinal outcomes, which may also be skewed (Fig. 1C).

Data distributions of study variables can have very different appearances that influence which statistical tests can be applied. Many statistical tests assume that the study data have a normal (bell-shaped or Gaussian) or near-normal distribution. If data do not have normal distribution, either the data should be transformed to obtain a near-normal distribution or other statistical tests that are appropriate for skewed distributions must be used (Petrie and Watson, 2013). When statistical tests are not appropriately applied, results can be nonsensical or misleading.

A common error in animal research is to use incorrect statistical tests for categorical data such as CIS or lameness scores, which are not quantitative and do not have a normal distribution. Appropriate methods to evaluate skewed, qualitative data compare the probability of each treatment or observation group having the observed percentages of animals in each category, given the assumption that all groups were drawn from the same population (Davis et al., 2009). Once the

appropriate tests are selected, the next step is to interpret what the inferential statistics are communicating relative to a clinical or management question.

RELATIONSHIP BETWEEN INDEPENDENT AND DEPENDENT VARIABLES

Because many studies in animal and veterinary science involve questions about the relationship between 2 or more variables, the statistical technique of regression analysis is frequently reported. Depending on the type of dependent variable and whether or not the model meets certain assumptions, different models (e.g., linear, logistic, Poisson, Cox, polynomial, ridge, etc.) are appropriate to analyze different study data (Dugard and Staines, 2010). Multiple linear regression is a commonly reported regression analysis, but this model makes several important assumptions, and if those assumptions are not met, the results may not be reliable (Kaps and Lambertson, 2004). In particular, violations of linearity or additivity or of independence are extremely serious and indicate that the linear regression model is not appropriate. Violations of homoscedasticity or normality of the error distribution can make the model results less trustworthy. Multiple variable prediction models for prognosis or diagnosis are particularly relevant to answer clinical or management questions in animal and veterinary science, and a reporting guideline has been published to address important aspects of model development and validation (Collins et al., 2015).

Internal and External Research Validity

Inferential statistics are not able to differentiate between bias and meaningful treatment differences, so experimental design features to control for bias and confounding must be assured before inferential statistics are used or interpreted. Internally valid studies collect data in a repeatable manner while controlling for bias and confounding through a research design to remove or control factors that could systematically influence study outcomes away from the truth. Studies with high internal validity will generate results influenced only by the study factors of interest, whereas the outcomes of studies with poor internal validity could be altered away from the truth due to factors that are not accounted for in the study design.

Statistical tests do not evaluate whether or not data are biased; therefore, studies with low internal validity should not be used for decision-making regardless of the statistical results (White and Larson, 2015a,b). If incorrect experimental design or statistical methods are used, recognize that the study results can be nonsense at best and misleading at worst. Unfortunately, readers cannot assume reports in animal science and veterinary journals are always based on appropriately designed studies and correctly applied statistical tests (Shott, 2011).

A number of reporting guidelines have been developed with the primary purpose being to assist authors to write accurate, complete, and clear reports of their research studies. In addition to their value to authors, these guidelines also aid readers critically appraise and interpret scientific literature by focusing attention on the aspects of materials and methods and study results that could indicate increased risk of bias in the study design, analysis, or reporting. The Equator (Enhancing the QUAlity and Transparency of health Research) network website (http://www.equator-network.org; accessed 20 Oct. 2015) provides an up-to-date source for reporting guidelines for many different study types (Simera and Altman, 2013).

External validity characterizes the appropriateness of applying the research results to populations other than the study population. For example, if a study was performed using a population of lightweight calves, the results should be extrapolated only to similar populations of lightweight calves if the biological system evaluated is likely to be different between weight groups.

Confirmatory versus Discovery Hypotheses

Research hypotheses can be placed into 2 broad categories: confirmatory and discovery. A confirmatory hypothesis tests a specific relationship that is proposed during the study design phase and is supported by previous investigations, whereas discovery hypotheses are suggested and tested based on the data generated by the study. Comparing the proportion of animals diagnosed with a particular disease with a positive clinical response when given a new therapy to the percentage of animals given a placebo control treatment that have a positive clinical response is an example of a confirmatory hypothesis if previous research identified the new therapy as a potentially effective treatment. In contrast, an example of a discovery hypothesis would be to investigate potential associations between multiple laboratory indices and animal disease status without identifying a specific association of interest before initiating the study.

Using statistical tests to evaluate multiple discovery hypotheses suggested by the data must be done cautiously to avoid mistaking chance statistical associations for biologically meaningful relationships. Many statistical tests report a *P*-value representing the probability that if the treatment or observation groups were truly not different, a difference as great as or greater than the one identified between 2 study variables is due to chance. If there are no true differences between treatment or observation groups, the likelihood of incorrectly concluding that that

a difference exists will increase in proportion with the number of discovery hypotheses examined.

Consider a study collecting samples from multiple animals to evaluate the relationship between complete blood count components with the presence or absence of infectious respiratory disease. We could use a confirmatory hypothesis stated before the study begins and based on previous research that total white cell count is associated with the presence of pneumonia (Ellis et al., 1998; Hanzlicek et al., 2010). If the inferential statistics support this relationship, our hypothesis is strengthened. Therefore, we feel more comfortable extrapolating clinical conclusions based on this data to similar populations.

In a contrasting study design, if our study did not have a specific confirmatory hypothesis but, rather, started with discovery hypotheses that 1 or more complete blood count components could be associated with animals having respiratory disease, our interpretation of the statistical results would be different compared with interpretation in the confirmatory study. If 20 independent blood components were evaluated with a *P*-value of ≤0.05 designated for statistical significance, then it is likely that even if no true biologic relationship exists between any of the tested variables and pneumonia status, at least 1 association with the outcome of interest will likely have a *P*-value less than 0.05 due to chance. Our inferences using the same data and the same significance level should change, based on the type of hypothesis and evaluation methods. Methods exist (e.g., Bonferroni correction, Tukey adjustment, and decreasing significance level) that allow inferences based on studies exploring multiple hypotheses and making multiple comparisons within an individual statistical model and should be used and reported in these types of studies.

Interpreting statistics should always be done with a clear understanding of the research hypothesis. A well-designed and well-conducted study with a single confirmatory hypothesis can often be used to strongly influence decision-making, whereas a study with many discovery hypotheses is best interpreted as a project to identify potential future confirmatory studies.

INTERPRETING *P*-VALUES BELOW THE SIGNIFICANCE LEVEL (E.G., $P \leq 0.05$ OR $P \leq 0.01$, ETC.)

Statistical tests determine the probability that if there were truly no difference between treatment or observation groups, a difference in outcomes as great as or greater than the one observed in the study could have been due to chance. The probability is commonly expressed as a *P*-value. Although *P*-values less than 0.05 have traditionally been considered "statistically significant," other significance (α) levels may be selected in different situations based on the type of research question and the level of certainty desired (Anderson et al., 1990). In response to criticisms and misunderstandings surrounding the interpretation of *P*-values, the American Statistical Association published a statement that, among other considerations, strongly stated that "Practices that reduce data analysis or scientific inference to mechanical 'bright-line' rules (such as $P < 0.05$) for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision-making. A conclusion does not immediately become 'true' on one side of the divide and 'false' on the other." And, "the widespread use of 'statistical significance' (generally interpreted as $P \leq 0.05$) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process" (Wasserstein and Lazar, 2016, p. 131).

Importantly, calculations of *P*-values do not take into consideration the risk for bias or confounding, meaning that a biased study may very well have a very low *P*-value but not provide true information on the effect of the treatment or risk factor of interest. Therefore, *P*-values provide useful information only in well-designed and well-conducted research studies and can be misleading if interpreted from studies that are poorly designed to control for bias and confounding. In addition, *P*-values do not provide information about the magnitude of difference or the importance of an effect because identical *P*-values calculated from studies with different sample size do not provide evidence for identical strength of association or strength of evidence (Gliner et al., 2002; Wagenmakers, 2007).

When interpreting the data from many well-designed and well-conducted studies, using a *P*-value of 0.05 or less to determine statistical significance results in a relatively low to moderate likelihood that one will conclude that a difference between treatments or risk factors is present when there is truly no difference. However, it is important to recognize that the *P*-value is not a direct estimate of the likelihood that the study findings are incorrect; rather, the risk of erroneous conclusions is related to both the *P*-value and the actual (but unknown) relationship between the variables of interest and the study outcome. This fact should cause one to interpret a *P*-value from a confirmatory study based on a strongly supported hypothesis very differently than the same *P*-value from a discovery study based on little or no prior supporting evidence.

Interpreting scientific literature to make clinical and management decisions is somewhat analogous to a clinical interpretation of a diagnostic test. Diagnostic test accuracy is often expressed in terms of sensitivity (the ability to correctly identify positive animals) and specificity (the ability to correctly identify negative animals). Although these variables are important, the likelihood

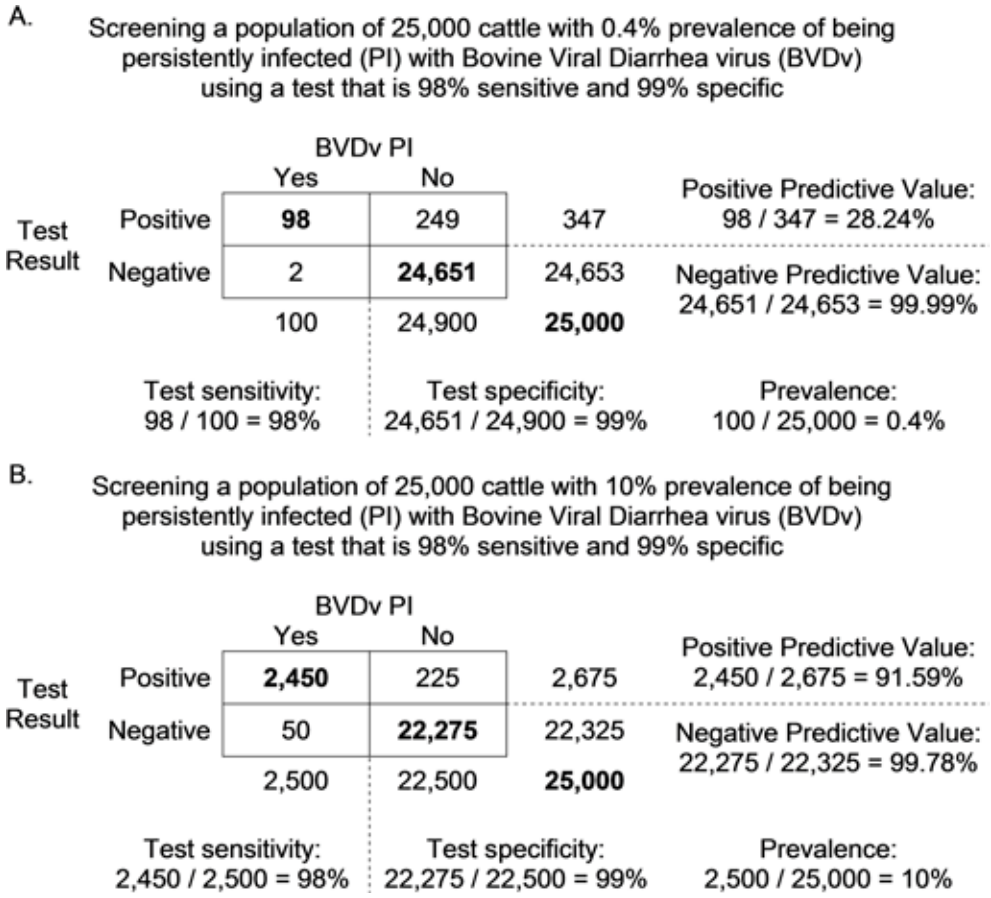

Figure 2. Diagnostic test illustration of the influence of the commonness of a characteristic (being persistently infected [PI] with bovine viral diarrhea virus [BVDv]) on the interpretation of test outcomes when the test characteristics (diagnostic sensitivity and specificity) remain constant. The interpretation of a positive test result (positive predictive value) is very different when the prevalence of the tested condition is low (A) compared with when the prevalence of the tested condition is high (B), even though the test characteristics do not change.

of false-positive and false-negative testing classification errors is strongly influenced by the true prevalence of the disease in the population. As an example, consider testing a population of cattle for bovine viral diarrhea virus (**BVDv**; Fig. 2). Diagnostic tests with sensitivity and specificity of 98 and 99%, respectively, are available (Nickell et al., 2011). The expected prevalence for cattle persistently infected with BVDv in the general population of feeder calves at the time of arrival at a feed yard is around 0.4% (Fulton et al., 2006). The calculated negative predictive value of using this test in this population is nearly 100% whereas the positive predictive value of using the test in this population is only 28% (Fig. 2A). In other words, a calf that we identify as test-positive is truly positive only 28% of the time. Despite high test specificity (99%), the positive predictive value (or the assurance that test-positive animals arriving at a feed yard are truly positive) is relatively low. Conversely, if we apply the same test in a herd that has been confirmed to have at least some cattle infected with BVDv and we expect the herd prevalence to be 10%, our positive predictive value becomes 92% (Fig. 2B). The sensitivity and specificity of the diagnostic test did not change, but our clinical interpretation is influenced by the underlying true prevalence within the population.

In a similar manner, the true (but not precisely known) relationship between the variables of interest and the study outcome impacts how we should interpret our statistical results. Similar to limited certainty about true disease prevalence in a given situation, we lack a precise estimate of how many of our studies are testing factors with a true difference in outcomes. However, we can place studies into broad categories such as discovery studies, where we expect that a low percentage of time there is truly a difference between observation groups or treatments (i.e., the null hypothesis is often true), and confirmatory work, where, based on previous studies, we expect a greater likelihood of finding a true biologic difference between observation groups or treatments. Providing an estimate of the percentage of studies that would truly result in differences if all knowledge was known allows us to generate calculations similar to the positive and negative predictive values. Although we might expect a relatively low risk of incorrect study interpretations (or times when we conclude there is truly a difference between treatment or

observation groups yet one does not exist) using a significance level of 0.05, it is important to recognize that most discovery studies are investigating observation group or treatment populations that are truly not different. Therefore, if novel treatments or risk factors investigated in discovery studies are truly different from controls only 10% of the time, we would expect, when using a significance level of 0.05, to classify differences as "statistically significant" that approximately 38% of those conclusions could be in error (Fig. 3A).

Occasionally in discovery studies, the significance level is raised as high as a *P*-value of 0.10 to allow greater exploration of possible associations; however, if a *P*-value of 0.10 is applied to the example in Fig. 3A, the risk of error in studies identified with a statistically significant difference between treatment or observation groups could be as high as 53%. Or, in other words, of all discovery studies that identified a statistically significant difference at $P < 0.10$, approximately half of those conclusions would be in error as no true difference exists. Therefore, a supposedly stringent significance level of 0.05 may not be strict enough in true discovery work (Sterne and Smith, 2001). This limitation to simplistic interpretations of a specific *P*-value being classified as "statistically significant" and, therefore, being considered to support rejection of the null hypothesis of no effect may help explain recent findings that illustrated that many (47/53; 89%) recent landmark publications in global cancer research could not be replicated in subsequent work (Cull et al., 2012).

Not all studies are conducted in pure discovery mode with the associated high risk of investigating risk factors or treatments not different from controls. In confirmatory research studies, if hypotheses that risk factors or treatments result in outcomes different from controls were correct 40% of the time, the risk of concluding that a difference exists when none truly does using a significance level of $P \leq 0.05$ is much lower than the same level applied to discovery studies (8%; Fig. 3B). If the hypotheses that treatment or observation groups are different were correct most of the time (>60%) in an area of study, the likelihood of incorrectly concluding a difference exists using a significance level of $P \leq 0.05$ is low (<4%); however, research on topics already well characterized is rarely valuable for advancing scientific knowledge. Therefore, our assumption is that most research is performed on hypotheses where the true association is unknown and the likelihood of true differences between treatments or observation groups is relatively low; if *P*-values are incorrectly interpreted as an estimate of the risk of Type I error (i.e., based on the statistical analysis, concluding a difference between treatment or observation groups exists when there is truly no difference), the role of chance differences will be underestimated. Because one does not know the actual frequency that study variables result in true outcome differences, using a conservative significance level ($P \leq 0.05$, $P \leq 0.01$, or $P \leq 0.001$) provides a more rigorous statistical test and improves robustness of findings compared with accepting *P*-values > 0.05 as being statistically significant. Knowledge of the biology and expected outcomes in an area of scientific investigation is needed to appropriately evaluate and interpret results (Sterne and Smith, 2001).

INTERPRETING *P*-VALUES ABOVE THE SIGNIFICANCE LEVEL (E.G., $P > 0.05$ OR $P > 0.01$, ETC.)

An inference test that returns a *P*-value greater than the selected significance level indicates there is not enough evidence to attribute the observed difference to the treatment or observation groups in the study; chance may have produced an as-great or greater observed difference (Greenland et al., 2016). Regardless of the magnitude of the *P*-value or the numerical difference between treatment or observation groups evaluated, a *P*-value above the selected significance level (e.g., $P > 0.05$) does not mean that the treatment or observation groups are equivalent (Greenland et al., 2016). Inferences concerning an outcome difference between treatment or observation groups with a calculated *P*-value of 0.15 should not be different from comparisons with a *P*-value of 0.95 based on the *P*-values alone because both values lead to the conclusion there is little evidence of a difference among the study groups. For almost all studies, the hypothesis centers around testing whether differences between treatment observation groups can be inferred to be due to the treatment rather than chance. If the statistical tests used in the study fail to remove chance as a reasonable explanation for observed differences (e.g., *P*-value > 0.05 or *P*-value > 0.01 depending on the a priori designated significance level), we do not know whether the study did not have sufficient replicates to identify a true difference or whether there is truly no effect of the treatment or observation group. A *P*-value above the significance level does not distinguish between these 2 explanations and no further conclusions should be drawn from the data. Post hoc power analysis methods can be used to determine the ability to detect a certain magnitude of difference between treatment or observation groups based on observed outcomes; however, the use of post hoc power analysis may be misleading and is generally discouraged (Smith and Bates, 1992; Goodman and Berlin, 1994).

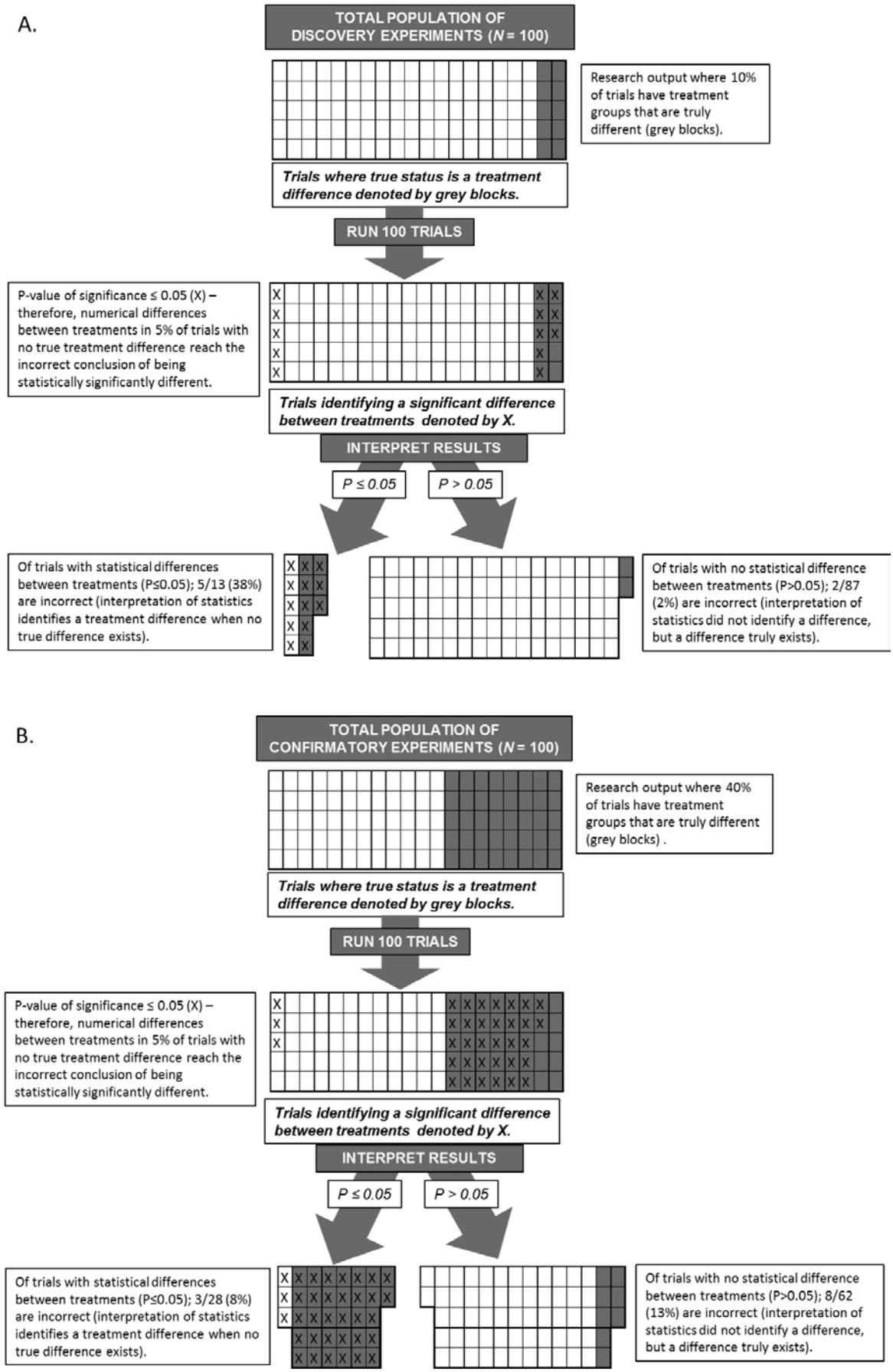

Figure 3. Depiction of studies (n = 100) to evaluate the expected results of discovery (A; 10% of trials have treatment or observation groups that are truly different) or confirmatory studies (B; 40% of trials have treatment or observation groups that are truly different) where trials identifying treatment or observation group differences are denoted with an "X" and the true state of the natural world is denoted by no difference between observation groups or treatments

Interpreting P*-Values in Light of Sample Size*

Statistical results should be interpreted in light of the overall study design and the sample size. Increasing the number of replicates in each treatment or observation group improves the accuracy of the estimate of the study outcomes, but the amount of replication within each study group is a compromise between improving accuracy of the estimate of the effect and the constraining cost and effort (Voisinet et al., 1997a). The power of a study to find a difference can be influenced by the study design, the type of outcome variable being measured (quantitative vs. qualitative), and the number of replicates (sample size; Voisinet et al., 1997b). Small sample sizes may overestimate population variability and thus make it challenging to identify true differences of small magnitude; therefore, when small studies show no statistically significant differences, the findings should not be overinterpreted because in reality, little additional information has been gained from the study.

Conversely, when a small study shows a significant finding, the magnitude of effect between treatment or observation groups is either truly large or, by chance, is uncharacteristically large in the specific study population. In other words, if 2 treatments or risk factors are truly different but the true magnitude of difference is small, an underpowered study will find a statistically significant difference only when the numerical estimates for each treatment or observation group were obtained from the opposite tails of the population distributions. The magnitude of effect exhibited in studies with small sample size that report statistically significant differences is likely greater than the true magnitude exhibited between treatment or observation groups in the larger population. In addition, a statistically significant difference identified with a small sample size may also challenge the appropriateness of extrapolating study results to other populations. The same number of animals (or other experimental units) used in the study may markedly differ from other samples that could be taken from the larger population, which may inhibit the ability to extrapolate results to a larger population (McGrath, 1987). If a study performed with 6 animals finds a significant difference among treatments, the findings are likely real but at the high end of the range of possible results obtained by sampling the population. In addition, it may be unlikely these 6 animals accurately represent all of the variability present in the whole population; therefore, the study may need to be repeated to improve the precision of the estimate and the external validity of the findings.

Interpreting Confidence Intervals

In addition to or in place of calculating *P*-values for a hypothesis test, CI can also be calculated to express the degree of uncertainty associated with the study outcome. Whereas a *P*-value only provides information about how incompatible the study data are with the null hypothesis (given that the study design and statistical methods are appropriate), a CI provides that information as well as information about the expected magnitude of the effect. Commonly reported CI are 90, 95, and 99%. For a 90% CI, if sampling from a population was repeated 100 times, the CI calculated for each sample population would contain the true population parameter of interest (e.g., mean, median, relative risk, odds ratio, proportion, etc.) 90 times and the true population parameter would fall outside the CI for each sample population 10 times. The CI provides a range of estimates that the study suggests contains the true population parameter; however, the CI does not indicate the probability of the true population parameter within the interval identified from the study data (Trafimow and Marks, 2015; Greenland et al., 2016).

Selection of the CI to report is related to the desired precision and confidence of the estimate, in that greater confidence can be achieved with wider, less precise CI whereas, in contrast, narrower CI provide a more precise estimate with less confidence that the population parameter lies within the CI. The 95% CI will have a greater width than the 90% CI, and the difference in width of the 2 intervals is due to the probability that the true population value will lie within the CI calculated from multiple samplings of the same population. Using a 90% CI increases the hazard of concluding that a numerical difference identified from a study sample reflects a true difference in the population parameter when, in fact, no difference is present (increased risk of Type I error) compared with using a 95% CI. A 99% CI results in the least risk of making a Type I error but the greatest risk of failing to reject the null hypothesis of no difference between treatment or observation groups based on the statistical analysis when a difference truly exists (increased risk of Type II error).

Within a selected CI (e.g., 95%), a wide CI communicates that there is a great deal of uncertainty associated with the sample-estimated outcome of interest. When used to infer whether or not a statistically significant difference is present, if the CI for the difference between treatment or observation groups excludes zero, the study data are considered to be incompatible with the null hypothesis. Another way CI provide more information than *P*-values is when a comparison fails to reach the significance threshold; a CI provides a plausible magnitude of treatment effect, whereas a *P*-value that is larger than the chosen significance level (e.g., $P > 0.05$) does not provide any information beyond the study failing to reject the null hypothesis.

Table 1. Glossary of commonly used statistical and research design terms

Term	Practical definition	Why it is important to researchers
Bias	Any factor that could systematically influence the study outcome away from the truth.	If study results are biased, inferences should not be made from the study.
Blinding	Ensuring that no one involved in describing or measuring study variables, care of the animals, or analysis of the data has knowledge of treatment exposure.	Blinding is one of the most effective mechanisms to prevent bias in outcome assessment and is essential when subjective variables are the primary outcomes.
Confidence interval (CI)	An estimation of the proportion of sample population intervals that would contain the true population a specified percent of the time. For a 95% CI, if a population was sampled 100 times, the 95% CI from each sample population would encompass the true population parameter 95% of the time. The larger the sample size, the more precision there is in the outcome estimated resulting in a smaller CI.	Similar to the *P*-value, a CI provides information about the uncertainty surrounding an estimate of the numerical difference in an outcome of interest between sampled treatment or observation groups as a reflection of the true difference between populations. In contrast to a *P*-value, a CI also provides information about the probable magnitude of effect, which is helpful when considering the clinical relevance of the results.
Confounding	A specific type of bias when a factor other than the treatment or observation factor of interest is associated with the study outcome but this factor is not evenly distributed between treatment or observation groups.	When confounding is present (or possible) and not controlled, distinguishing treatment effects from the effect of the potential confounder is impossible, leading to an inability to draw firm conclusions from the research.
Experimental or observational unit	The smallest independent physical unit that is assigned to a treatment (experimental study) or observed (observational study), and each experimental or observational unit must be able to receive a different treatment.	The experimental or observational unit for each hypothesis must be correctly identified by the investigator to ensure that the study has adequate sample size (e.g., power) and that the statistics tests were properly performed. Livestock studies often have hierarchical data with animals nested within pens nested within buildings nested within farms or with repeated measurements of the same units being taken over time. These types of hierarchical data structures makes selection of the correct experimental or observation unit and the appropriate statistical test more challenging and may result in different experimental or observational units being appropriate for different hypothesis in the same study.
External validity	The ability for study results to accurately be generalized to other populations.	Research may be internally valid yet performed in a population much different from the population of clinical interest; therefore, extrapolating study results may not be possible.
Internal validity	The study design is appropriate for the hypotheses and study variables while controlling for potential issues related to bias.	If the study is not internally valid, conclusions based on results should not be made.
Interaction	The effect of one variable on the outcome of interest is modified by the effect of another variable.	Interactions are relatively common in biologic studies, and if present, they can influence interpretation of study outcomes.
Least squares mean	The result of a statistical analysis to approximate the solution in a model fitting the outcome and adjusting for other variables in the model.	Least squares means are calculated from a model that adjusts the estimated mean based on variables included in the model. This adjustment should result in a more accurate estimate of the population mean than a simple arithmetic mean of the sample data.
Multivariable analysis	A statistical analysis that incorporates the relationship of more than 1 variable when evaluating the outcome of interest.	Biologic systems are complex, and often, experimental or observational units are not completely independent. Multivariable analyses allow for evaluation of effects while adjusting for potential variables that may be confounding, resulting in more accurate estimate of effects.
Null hypothesis	The starting assumption for most research; the assumption that there is no difference among treatment or observation groups.	Because the initial assumption is no difference between treatment or observation groups, if the statistical tests fail to identify a difference, no real conclusions should be drawn from the findings (one can say only that the treatments were not statistically different at the magnitude observed with this study sample size). Failure to disprove the null hypothesis does not indicate that treatment or observation groups are the same, only that they did not statistically differ in this study.
Numerical differences	The outcomes (e.g., mean, median, relative risk, odds ratio, proportion, etc.) of 2 treatment or observation groups differ, but the difference could be due to chance, bias, or true treatment or observation group differences.	Study findings may be described as numerically different but the difference could be due solely to chance and biological variability. If a statistical difference was not identified in the presence of numerical differences between treatment or observation groups, this means either the sample size was too small to detect a true population difference of small magnitude or no difference exists. Conclusions should not be based on numerical differences alone.

Continued

Table 1. (cont.)

Term	Practical definition	Why it is important to researchers
P-value	The result of a statistical test that reports the probability of an outcome difference as great as or greater than that described by the study data being incompatible with a specific model for the data (the model typically assumes that study populations are truly not different as well as making other assumptions).	The *P*-value is used to determine statistical significance but can be interpreted only if the study is internally valid and the interpreter has knowledge about the study design and the topic being investigated. A *P*-value does not test for bias, study design, or the appropriateness of applying study results to other populations dissimilar to the population tested. In addition, it does not measure the probability that the hypothesis is true or the probability the data were produced by chance.
Pseudoreplication	The error of treating multiple observations from the same experimental or observational unit as replications of independent experimental or observational units.	Taking multiple samples from a single experimental or observational unit and treating them as independent samples can lead to the danger of concluding that a difference exists when that may not be true.
Randomization	A common process of assigning experimental units to treatment groups used to prevent inadvertent bias based on selection criteria.	Randomization is the primary mechanism to mitigate the danger of selection bias and confounding. Randomization attempts to prevent a factor outside the study criteria being present in unequal distributions among the treatment groups.
Sample size	The number of experimental or observational units for each treatment or observation factor group within the study.	Adequate sample size is based on the outcome of interest, the expected variability of the outcome, and the expected magnitude of effect of the treatment or observation factor on the outcome. Without adequate sample size, studies could be referred to as underpowered and are unlikely to identify true differences.
Statistically significant	The results of a statistical test to compare results. The specific definition of the threshold of "significance" may vary among researchers, but significance (α) levels of $P \leq 0.05$ and $P \leq 0.01$ are commonly used.	Denoting statistical significance describes the probability of observing a difference as large as or larger than that identified in the study with the current sample size if there were truly no differences between treatment or observation groups (see *P*-value). This designation does not differentiate between studies with and without true differences between study groups, nor does a statistically significant difference indicate a finding that is necessarily biologically meaningful, and findings should be interpreted accordingly. The α selected to be considered statistically significant should be influenced by the number of comparisons being made, the type of hypothesis (confirmatory vs. discovery) being tested, and expert knowledge of the study subject.
Type I error	Based on the statistical analysis, concluding that a difference among treatment or observation groups exists when there is truly no difference.	Even when treatment or observation groups do not truly differ, individual studies will identify statistical differences among treatment or observation groups. The frequency with which this error occurs compared with true parameter differences is based on the significance (α) level or the CI selected and the underlying probability that the treatment or observation groups are truly different (discovery vs. confirmatory research).
Type II error	Based on the statistical analysis, failing to reject the null hypothesis of no difference between treatment or observation groups when a difference truly exists.	Even when treatment or observation groups truly differ**[**, individual studies can fail to identify statistical differences among study groups. The frequency with which this error occurs is based on the power of the study (influenced by sample size and the selected beta or CI).

Biological versus Statistical Significance

The reader of a scientific paper must determine how much difference between treatment or observation groups is meaningful enough to support application of the findings to specific animal populations. The magnitude of the difference between treatment or observation groups that reach the a priori test of being statistically significantly different is influenced by sample size (number of replicates in each treatment or observation group) and the inherent variability within the outcome being measured. Typically, when a small study identifies a statistically significant difference, the difference is relatively large, whereas a very large study can detect small numerical differences between treatment or observation groups.

Statistical tests do not provide information about the importance of numerical differences between treatment or observation groups. The reader must determine if the amount of difference between treatment or observation groups is biologically meaningful and represents a logical finding. A project evaluating physiological changes before and after bacterial respiratory disease challenge in cattle reported several statistical differences in the hematologic profile (Hanzlicek et al., 2010). In the study, there were significantly fewer leukocytes before challenge (9.58×10^3 cells/μL) compared with the day after

challenge (12.57×10^3 cells/μL; Hanzlicek et al., 2010). The findings were statistically significant, yet both findings are located in the middle of the normal laboratory reference range for leukocytes (7.0 to 14.0×10^3 cells/μL). These results represent a logical biological process (increased white blood cells following induced bacterial respiratory disease) and a statistical difference was identified; however, a difference of this magnitude is not biologically meaningful if the clinical objective is to identify a measure that could be used to rapidly diagnose respiratory disease. Care should be taken to interpret not just if there is a difference between treatment or observation groups but whether or not the difference would influence a clinical or management decision (Gliner et al., 2002).

The interpretation of numerical differences between treatments is not always easy, nor is it intuitive. There are instances when the temptation is to modify ones interpretation to reflect inherent biases, allowing numerical differences that do not reach a priori established significance levels or CI tests of statistical significance to unduly influence clinical or management decisions. Knowledge about previous research is needed to determine whether a specific study design was discovery or confirmatory. If previous research supports the null hypothesis of no difference between study treatment or observation groups, then a numerical difference in a well-powered study that does not reach the significance level may add evidence of no effect. However, if no previous research has been performed comparing the study treatment or observation groups, then an initial study most likely serves a discovery role and additional research is needed to appropriately evaluate outcomes. Although inferential statistics can give a qualitative indication of whether or not the study observations are incompatible with the null hypothesis (below the significance level), if the null hypothesis of no treatment or observation group effect cannot be rejected, numerical differences alone should not be used to drive clinical and management decisions without information about study design, biology, and previous research (Wasserstein and Lazar, 2016).

SUMMARY AND CONCLUSIONS

A basic understanding of statistical principles is useful when interpreting studies to enhance decision-making. Statistics provide a framework to transfer information from relatively small, well-designed research studies into information that can be applied to broader populations to support clinical and management decisions. Table 1 provides a glossary of common statistical terms encountered in animal research as well as these terms' practical importance for readers. It is important to realize that statistical results should be interpreted in light of a specific clinical or management question as well as content-specific expertise and knowledge about biology and study design to best use inferential statistics to support clinical and management decisions.

LITERATURE CITED

Anderson, K. L., C. A. Neff-Davis, L. E. Davis, and V. D. Bass. 1990. Pharmacokinetics of flunixin meglumine in lactating cattle after single and multiple intramuscular and intravenous administrations. Am. J. Vet. Res. 51:1464–1467.

Collins, G. S., J. B. Reitsma, D. G. Altman, and K. G. M. Moons. 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Ann. Intern. Med. 162:55–63. doi:10.7326/M14-0697

Cull, C. A., Z. D. Paddock, T. G. Nagaraja, N. M. Bello, A. H. Babcock, and D. G. Renter. 2012. Efficacy of a vaccine and a direct-fed microbial against fecal shedding of *Escherichia coli* O157:H7 in a randomized pen-level field trial of commercial feedlot cattle. Vaccine 30:6210–6215. doi:10.1016/j.vaccine.2012.05.080

Davis, J. L., G. W. Smith, R. E. Baynes, L. A. Tell, A. I. Webb, and J. E. Riviere. 2009. Update on drugs prohibited from extralabel use in food animals. J. Am. Vet. Med. Assoc. 235:528–534. doi:10.2460/javma.235.5.528

Dugard, P. J. T., and H. Staines. 2010. Approaching multivariate analysis: A practical introduction. Routledge, Hove, UK, and New York, NY.

Ellis, J. A., K. H. West, V. S. Cortese, S. L. Myers, S. Carman, K. M. Martin, and D. M. Haines. 1998. Lesions and distribution of viral antigen following an experimental infection of young seronegative calves with virulent bovine virus diarrhea virus-type II. Can. J. Vet. Res. 62:161–169.

Fulton, R. W., B. Hessman, B. J. Johnson, J. F. Ridpath, J. T. Saliki, L. J. Burge, D. Sjeklocha, A. W. Confer, R. A. Funk, and M. E. Payton. 2006. Evaluation of diagnostic tests used for detection of bovine viral diarrhea virus and prevalence of subtypes 1a, 1b, and 2a in persistently infected cattle entering a feedlot. J. Am. Vet. Med. Assoc. 228:578–584. doi:10.2460/javma.228.4.578

Gliner, J. A., N. L. Leech, and G. A. Morgan. 2002. Problems with null hypothesis significance testing (NHST): What do the textbooks say? J. Exp. Educ. 71:83–92. doi:10.1080/00220970209602058

Goodman, S. N., and J. A. Berlin. 1994. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann. Intern. Med. 121:200–206. doi:10.7326/0003-4819-121-3-199408010-00008

Greenland, S., S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman. 2016. Statistical tests, *P*-values, confidence intervals, and power: A guide to misinterpretations. Eur. J. Epidemiol. 31:337–350. doi:10.1007/s10654-016-0149-3

Hanzlicek, G. A., B. J. White, D. Mosier, D. G. Renter, and D. E. Anderson. 2010. Serial evaluation of physiologic, pathological, and behavioral changes related to disease progression of experimentally induced *Mannheimia haemolytica* pneumonia in postweaned calves. Am. J. Vet. Res. 71:359–369. doi:10.2460/ajvr.71.3.359

Kaps, M., and W. R. Lambertson. 2004. Multiple linear regression. In: M. Kaps and W. R. Lambertson, editors, Biostatistics for animal science. CABI Publishing, Wallingford, UK. p. 154–184. doi:10.1079/9780851998206.0154

Kottner, J., L. Audige, S. Brorson, A. Donner, B. J. Gajewski, A. Hrobjartsson, C. Robers, M. Shoukri, and D. L. Streiner. 2011. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J. Clin. Epidemiol. 64:96–106. doi:10.1016/j.jclinepi.2010.03.002

McGrath, P. A. 1987. An assessment of children's pain: A review of behavioral, physiological and direct scaling techniques. Pain 31:147–176. doi:10.1016/0304-3959(87)90033-9

Nickell, J. S., B. J. White, R. L. Larson, D. G. Renter, and M. W. Sanderson. 2011. A simulation model to quantify the value of implementing whole-herd bovine viral diarrhea virus testing strategies in beef cow-calf herds. J. Vet. Diagn. Invest. 23:194–205. doi:10.1177/104063871102300202

Perino, L. J., and M. Apley. 1998. Clinical trial design in feedlots. Vet. Clin. North Am. Food Anim. Pract. 14:343–365. doi:10.1016/S0749-0720(15)30258-9

Petrie, A., and P. Watson. 2013. Further aspects of design and analysis. In: Statistics for veterinary and animal science. 3rd ed. Wiley-Blackwell, West Sussex, UK. p. 181–183.

Schlageter-Tello, A., E. A. M. Bokkers, P. W. G. Groot Koerkamp, T. VanHertem, S. Viazzi, C. E. B. Romanini, I. Halachmi, C. Bahr, D. Berckmans, and K. Lokhorst. 2014. Effect of merging levels of locomotion scores for dairy cows on intra- and interrater reliability and agreement. J. Dairy Sci. 97:5533–5542. doi:10.3168/jds.2014-8129

Shott, S. 2011. Detecting statistical errors in veterinary research. J. Am. Vet. Med. Assoc. 238:305–308. doi:10.2460/javma.238.3.305

Simera, I., and D. G. Altman. 2013. Reporting medical research. Int. J. Clin. Pract. 67:710–716. doi:10.1111/ijcp.12168

Smith, A. H., and M. N. Bates. 1992. Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. Epidemiology 3:449–452. doi:10.1097/00001648-199209000-00011

Sterne, J. A., and D. G. Smith. 2001. Sifting the evidence – What's wrong with significance tests? BMJ 322:226–231. doi:10.1136/bmj.322.7280.226

Theurer, M. E., D. G. Renter, and B. J. White. 2015. Using feedlot operational data to make valid conclusions for improving health management. Vet. Clin. North Am. Food Anim. Pract. 31:495–508. doi:10.1016/j.cvfa.2015.05.004

Trafimow, D., and M. Marks. 2015. Editorial. Basic Appl. Soc. Psych. 37:1–2. doi:10.1080/01973533.2015.1012991

Voisinet, B. D., T. Grandin, S. F. O'Connor, J. D. Tatum, and M. J. Deesing. 1997a. *Bos indicus*-cross feedlot cattle with excitable temperaments have tougher meat and a higher incidence of borderline dark cutters. Meat Sci. 46:367–377. doi:10.1016/S0309-1740(97)00031-4

Voisinet, B. D., T. Grandin, J. D. Tatum, S. F. O'Connor, and J. J. Struthers. 1997b. Feedlot cattle with calm temperaments have higher average daily gains than cattle with excitable temperaments. J. Anim. Sci. 75:892–896. doi:10.2527/1997.754892x

Wagenmakers, E. J. 2007. A practical solution to the pervasive problems of *p* values. Psychon. Bull. Rev. 14:779–804. doi:10.3758/BF03194105

Wasserstein, R. L., and N. A. Lazar. 2016. The ASA's statement on *p*-values: Context, process, and purpose. Am. Stat. 70:129–133. doi:10.1080/00031305.2016.1154108

White, B. J., and R. L. Larson. 2015a. Systematic evaluation of scientific research for appropriateness of data analysis to improve clinical decision making. J. Am. Vet. Med. Assoc. 247:759–762. doi:10.2460/javma.247.7.759

White, B. J., and R. L. Larson. 2015b. Systematic evaluation of scientific research for clinical relevance and control of bias to improve clinical decision making. J. Am. Vet. Med. Assoc. 247:496–500. doi:10.2460/javma.247.5.496