Enhancing Clinical Decision-Making: Challenges of making decisions on the basis of significant statistical associations

Loic Desquilbet

Desquilbet L. Enhancing Clinical Decision-Making: Challenges of making decisions on the basis of significant statistical associations. J Am Vet Med Assoc. 2020 Jan 15;256(2):187-193. doi: 10.2460/javma.256.2.187

Challenges of making decisions on the basis of significant statistical associations

Author: Loic Desquilbet, PhD

Professional affiliations:

1) Biostatistics and Clinical Epidemiology Service, Ecole Nationale Vétérinaire d'Alfort, UPEC, Maisons-Alfort, F-94700, France

2) U955 - IMRB, Inserm, Ecole Nationale Vétérinaire d'Alfort, UPEC, Maisons-Alfort, F-94700, France

Email addresses: loic.desquilbet@vet-alfort.fr, loic.desquilbet@gmail.com

A crisis of the reproducibility of studies reported in leading science journals emerged a few years ago[1-6]. One of the multiple origins of this crisis is that many statistically significant results obtained in some studies were not replicated or supported by other ones[7-9]. One of the reasons given is that many published results are false positives: authors wrongly generalize (i.e., infer) their results to the target population after obtaining a statistically significant result in their study sample[10,11]. P-hacking and HARKing are ones of the inappropriate methods of analyzing and interpreting study results, leading to false positive results. P-hacking (or "p-fishing") occurs when researchers collect data without a predetermined sample size, select data without *a priori* identification of inclusion and exclusion criteria, or select statistical analyses until non-significant results become significant[12-15]. HARKing occurs when researchers present a *post hoc* hypothesis based on their results as if it were an *a priori* hypothesis[16]. However, even in the absence of p-hacking, HARKing, association biases[17], or any other errors in scientific reporting, the probability to wrongly infer statistically significant results to the target population is very high is some situations. The purpose of the paper is to help researchers in veterinary science who do clinical research as well as clinicians interpreting those research findings appreciate the degree of (un)certainty when generalizing the study results to the target population of studied animals, according to the characteristics of the clinical study. This appreciation is also necessary to practice evidence-based veterinary medicine[18], especially when critically appraising evidence within the more general framework of the clinical decision-making process[19,20].

I. GENERAL CONCEPTS

Throughout this paper, the context will always be the following: researchers seek to provide evidence that there is a true association between one exposure (e.g., neutering status, being treated with a treatment *versus* placebo, surgical *versus* medical intervention) and one outcome

(e.g., disease occurrence, tumor remission, all-cause death). The statistical tests mentioned in this paper will therefore test such associations. The conclusions drawn from other statistical tests (e.g., those testing the normality of one distribution) will not be addressed. Furthermore, to facilitate the reading of the paper, the expression "significant association" will be used for "association classified as being statistically significant based on the p-value ($p \leq \alpha$)"; such expression does therefore not refer to the *clinical* significance of the association[21,22].

II. HYPOTHETICAL STUDIES

Two hypothetical studies will illustrate the concepts presented in the paper. For simplification purpose, these studies will be considered as feasible and ethical.

The context of the first study (study #1) is the following: Mullin et al. conducted a non-randomized study to assess the association between doxorubicin chemotherapy use (*versus* no therapy use) and death occurrence in dogs with presumptive cardiac hemangiosarcoma[23]. The statistically significant difference in death occurrence between the two groups suggested a potential effect of doxorubicin chemotherapy on time to death. In this context, the investigators of study #1 designed a randomized clinical trial to confirm the beneficial effect of doxorubicin chemotherapy within the first four months of use. To do so, they use the figures provided by the Kaplan-Meier curves in the paper of Mullin et al.[23] (45% and 5% of alive dogs at four months, respectively) in order to calculate the sample size with 80% of statistical power (n=79 dogs in each group). Then they followed the dogs during four months and compared the two groups of dogs on death occurrence by using the Kaplan-Meier method and the log-rank test. Study #1 can be considered as "confirmatory" since it is conducted to *confirm* the result of the previous study of Mullin et al.

The context of the second study (study #2) is the following: it has been suggested that masitinib monotherapy use has promising potential in treating canine epitheliotropic T-cell lymphoma[24].

In this context and based on these premises, the investigators of study #2 conducted a randomized clinical trial to assess the effect of masitinib in dogs diagnosed with multicentric lymphoma on (partial or complete) remission of their lymphoma. To do so, they randomly allocate 80 dogs with multicentric lymphoma in one out of two groups: one group receiving masitinib plus prednisone (n=40), the other one receiving prednisone only (n=40). This number of 80 dogs was not based on an *a priori* sample size calculation, but based on the available time to recruit dogs within in a predefined period. Then they followed the dogs during three months and compared the two groups on the presence of partial/complete lymphoma remission at three months. Study #2 can be considered as "exploratory" since it is the first one studying such association between masitinib plus prednisone use (*versus* prednisone use only) and lymphoma remission in such population of dogs with multicentric lymphoma.

III. REVIEW OF STATISTICAL CONCEPTS

Some statistical reminders about the null hypothesis, type-I and type-II errors, and statistical power are briefly provided below. These points are covered in more detail elsewhere[25].

A. The null-hypothesis and its acceptance or rejection

A statistical test testing the association between one exposure and one outcome is based on a "null-hypothesis" which is the absence of such association in a predefined (target) population[26]. For instance, the null-hypothesis of the log-rank test performed in study #1 is: "there is no association between doxorubicin chemotherapy use (*versus* no therapy use) and time to death in dogs with presumptive cardiac hemangiosarcoma.". If the null-hypothesis is rejected, one concludes that the study provides evidence supporting that there is a true association in the population between the exposure and the outcome. Most of the time, the rejection or acceptance of the null-hypothesis is based on the p-value provided by the statistical test: if the p-value is

less than or equal to a threshold value (α), the association is classified as "significant" in the study sample, and it is concluded that the null-hypothesis is false (rejection of the null-hypothesis). (Such use of the p-value and the "significant" / "non significant" approach has however contributed to the reproducibility crisis and is questioned by many scientists[27-30].)

B. Type-I and type-II errors and statistical power

When there is no true association between the exposure and the outcome in the population (or "when the null-hypothesis is true"), the probability of obtaining a significant association ($p \leq \alpha$) in the study sample is α (also known as the probability of type-I error).

When there is a true association between the exposure and the outcome in the population (or "when the null-hypothesis is false"), the probability of not obtaining a significant association ($p > \alpha$) in the study sample is equal to β (also known as the probability of type-II error). Therefore, is such situation, the probability of obtaining a significant association is (1-β), which is the value of the statistical power of the study.

IV. REVIEW OF DIAGNOSTIC TEST CONCEPTS

The sensitivity (Se) of a diagnostic test is the probability of a positive test result when the disease (or any health-related condition) is present, and the specificity (Sp) is the probability of a negative test result when the disease is not present. The positive predictive value (PPV) of a diagnostic test is the probability that an animal with a positive test result would have the disease. In a sample of N animals, Se, Sp, and PPV can be estimated by calculating the proportion of true-positive animals among diseased animals (Se), the proportion of true-negative animals among disease-free animals (Sp), and the proportion of true positive animals among test-positive animals (PPV).

From the frequencies of Table 1, Se, Sp and PPV can be expressed as:

$$Se = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{FP + TN}$$

$$PPV = \frac{TP}{FP + TP} \quad (1)$$

Table 1 can be re-expressed by using proportions instead of frequencies. To do so, let π be the prevalence rate of the disease in the population; it is therefore the probability of having the disease for an animal randomly drawn from this population before the result of the diagnostic test was known. If one randomly draws a sample of N animals from the population in which the prevalence rate is π, there will be N.π diseased animals and N.(1-π) disease-free animals. Among the N.π diseased animals, there will be N.π.Se true-positive animals; among the N.(1-π) disease-free animals, there will be N.(1-π).Sp true-negative animals. Table 2 presents the frequencies of Table 1 by using Se, Sp, and π in a sample of N animals.

Using Formula (1) of the PPV by replacing TP with N.π.Se and by replacing FP with N.(1-π).(1-Sp), one obtains:

$$PPV = \frac{\pi.Se}{(1-\pi).(1-Sp) + \pi.Se} \quad (2)$$

The Appendix illustrates the equivalence between Formula (1) and Formula (2) by using data from a hypothetical study. Therefore, in a situation where the prevalence rate of the disease (π) is low (for instance, 1%), a diagnostic test which has a very good Se (for instance, 80%) and an excellent Sp (for instance, 95%) will have a poor PPV (e.g., 14% provided by Formula (2) by using the previous values of π, Se, and Sp; see Appendix for illustrative data). A PPV of 14%

would be interpreted as: only 14% of animals with a positive test for the studied disease actually have the disease, while 86% of animals with a positive test are actually disease-free.

V. APPLICATION OF DIAGNOSTIC TEST CONCEPTS TO STATISTICAL TESTS

Although the interpretation of diagnostic tests and the interpretation of statistical tests have some important divergent concepts, some aspects of diagnostic test interpretation can serve as a reasonable metaphor for interpreting statistical tests[31]. To apply the concept of sensitivity, specificity, and positive predictive value of a diagnostic test to a statistical test testing the association between one exposure and one outcome, one must define what are the analogous terms for having the disease, for being disease-free, and for having positive and negative test results. With a statistical test, researchers seek for evidence that there is a true association in the population (i.e., evidence supporting that the null-hypothesis is false). A significant association in the study sample is in favor of the evidence they are seeking. In the situation of diagnostic tests, clinicians seek for evidence that the animal has the studied disease. A positive test for this disease is in favor of the evidence they are seeking. Therefore, the analogies described in Table 3 can be made[32].

The sensitivity of a diagnostic test is the probability of obtaining a positive test result when the disease is present. By analogy (see Table 3), the sensitivity of a statistical test is therefore the probability of obtaining a significant association when the null-hypothesis is false. This is analogous to the statistical power of a study (1-β). The specificity of a diagnostic test is the probability of obtaining a negative test result when the disease is absent. By analogy (see Table 3), the specificity of a statistical test is the probability of obtaining a non-significant association when the null-hypothesis is true. Since α is the probability of obtaining a significant association when the null-hypothesis is true, by analogy, the specificity of a statistical test is therefore (1-α).

VI. RELEVANCE OF CALCULATING THE PPV OF A STATISTICAL TEST

In practice, a clinician is much more interested in the PPV of diagnostic tests than the sensitivity and specificity of such tests. When a positive test result is obtained for an animal, one would like to know the probability that the animal has the studied disease (i.e., the PPV for having the disease). By analogy with statistical tests (see Table 3), when a significant association is obtained, one would like to know the probability that there is a true association in the population (i.e., the PPV for the null-hypothesis being false). For instance, the PPV of the log-rank statistical test performed in study #1 is the probability that there is a true association between doxorubicin chemotherapy use (*versus* no therapy use) and time to death in dogs with presumptive cardiac hemangiosarcoma, if the investigators obtained a significant association in their study sample.

The analogy for a low PPV of a statistical test indicates that the probability that there is a true association in the population is low even when the association is significant. In such situations of low PPV, it would not be surprising that a significant result in one study is difficult to replicate as being significant in another study of the same association targeting the same population[10].

To calculate the PPV of a statistical test, Formula (2) will be used by replacing the sensitivity, specificity, and π for diagnostic tests by their analogous values for statistical tests. For a statistical test, we previously determined that Se=(1-β) and Sp=(1-α). We must now interpret the value of π for statistical tests.

When researchers plan to design a study to test an association between one exposure and one outcome, they have some level of uncertainty that this association actually exists in the population. They must have such level of uncertainty because, if they knew with 100% certainty that the null-hypothesis is false, a study would not be necessary. Before conducting a study,

researchers therefore have in mind an *a priori* probability that the null-hypothesis is false in the studied population, which lies between 0 excluded and 1 excluded. For instance, because the association between doxorubicin chemotherapy use (*versus* no therapy use) and time to death in dogs with presumptive cardiac hemangiosarcoma has been previously suggested, the investigators of study #1 should have a higher level of certainty that this association truly exists, compared to the level of certainty of the investigators of study #2, where the association tested in study #2 has never been studied before.

In the context of diagnostic tests, π was the probability of having the disease for an animal randomly drawn from a population before the result of the test was known. By using the analogies presented in Table 3, for a statistical test, π would be the *a priori* probability that the null-hypothesis is false (i.e., the *a priori* probability that the association truly exists in the population). The "*a priori*" expression means "before the result of the statistical test is obtained from the study". (This expression refers to the notion of "prior information" in Bayesian statistics[33], in the context of the PPV of statistical tests[32,34].) As Browner and Newman wrote, the value of π for statistical tests is based on "biologic plausibility, previous experience with similar hypotheses, and knowledge of alternative scientific explanations"[32]. Therefore, in an exploratory study where researchers are the first ones to assess an association between one exposure and one outcome in a specific target population, it must be admitted that the *a priori* probability that such association truly exists (π) is low, despite the potential strong pathophysiological basis for this exploratory study[10].

We can rewrite Formula (2) of the PPV of a diagnostic test by replacing Se and Sp by their analogous value for a statistical test ((1-α) and (1-β), respectively):

$$PPV = \frac{\pi.(1-\beta)}{(1-\pi).\alpha + \pi.(1-\beta)} \quad (3)$$

With π being the *a priori* probability that the null-hypothesis is false (i.e., the probability that the association truly exists), α the type-I error, β the type-II error, and $(1-\beta)$ the statistical power.

The interpretation of Formula (3) is as follows. Suppose that the statistical power of study #2 is 80% when testing the association between masitinib plus prednisone use (*versus* prednisone use only) and lymphoma remission in dogs with multicentric lymphoma, and α is set at 5% (α=0.05). Suppose that the probability that this association truly exists is 1% (i.e., π=0.01); this low value of π can be explained by the fact that study #2 is exploratory and therefore involves much *a priori* uncertainty about the existence of such association. Based on the characteristics of study #2 (a statistical power of 80% and its exploratory status with π=0.01), the PPV calculated by using Formula (3) is 0.14 (14%). This value of 14% means that if the investigators of study #2 conduct this study and obtain a significant association, the probability that this association truly exists in the population of dogs with multicentric lymphoma is (only) 14%.

VII. FALSE POSITIVE REPORT PROBABILITY OF A STATISTICAL TEST

In the context where researchers would like to estimate the probability of wrongly concluding that there is a true association in the population after obtaining a significant one in the study sample, the complement of the PPV (1-PPV) is the most relevant indicator. This complement of PPV is called "false positive report probability" (FPRP)[35,36]. The FPRP is therefore the probability that there is no true association in the population after obtaining a significant association in the study sample. In other words, the FPRP quantifies the probability of wrongly concluding that there is a true association in the population after obtaining a significant one in the study sample. For instance, the FPRP of the log-rank statistical test used in study #1 is the probability of wrongly concluding that there is an association between doxorubicin chemotherapy use (*versus* no therapy use) and time to death in dogs with presumptive cardiac hemangiosarcoma after obtaining a significant association in the study sample.

We obtain the expression of the FPRP after calculation from Formula (3):

$$FPRP = 1 - PPV = \frac{(1-\pi) \times \alpha}{(1-\pi) \times \alpha + \pi \times (1-\beta)} \quad (4)$$

With π being the *a priori* probability that the null-hypothesis is false (i.e., the probability that the association truly exists), α the type-I error, β the type-II error, and (1-β) the statistical power. Figure 1 provides numerical examples of FPRP values according to selected values of π and (1-β).

VIII. MISINTERPRETATION OF THE TYPE-I ERROR AND P-VALUE

In the vast majority of cases, the Type-I error α is set at 5% (α=0.05), which is considered as a low value. Many researchers wrongly think that since α is low, the conclusion following a significant association is accompanied by a (same) low probability of error[37,38]. Similarly, p-values are commonly misinterpreted as the observed probability to wrongly reject the null-hypothesis, which means that researchers commonly and mistakenly interpret the p-value as if it were the FPRP[39]. For instance, if the p-value obtained in study #2 is 0.03, the investigators of study #2 would probably conclude with a mistaken belief that there is strong evidence for a true association between doxorubicin chemotherapy use (*versus* no therapy use) and time to death in dogs with presumptive cardiac hemangiosarcoma, with a 3% risk of error[28,40,41]. The p-value is actually the probability of the observed or more extreme results, if the null-hypothesis were true and if there were no bias when estimating the association. (The p-value has therefore no meaningful interpretation *per se* since its value is conditional on an hypothesis that nobody would know with 100% certainty whether it is true or false[42].)

IX. FACTORS CONTRIBUTING TO A HIGH FALSE POSITIVE REPORT PROBABILITY

The Formula (4) indicates that the FPRP of a statistical test performed in a study whose objective is to provide evidence that there is a true association between one exposure and one outcome depends on the probability of type-I error (α), on the statistical power of the study ($1-\beta$), and on the *a priori* probability that the null-hypothesis is false (π).

A. Impact of the value of the type-I error (α)

Suppose that, in study #1 designed with 80% of statistical power, the *a priori* probability that the null-hypothesis is false is 20% (i.e., $\pi=0.20$). If α is set to 1% ($\alpha=0.01$), the FPRP calculated from Formula (4) is 5%; if α is set to 5% ($\alpha=0.05$), the FPRP increases to 20%. Therefore, and more generally, the higher the type-I error (α), the higher the FPRP. This point is one of the origins of a scientific movement that questions the type-I error threshold of 5% ($\alpha=0.05$), and proposes to lower it to 0.5% ($\alpha=0.005$)[43]. This movement is however not shared by all scientists[44], and the convention for the type-I error (α) threshold set at 5% is likely to persist for years. From now on throughout the paper, the value of α will be set at 5% ($\alpha=0.05$).

B. Impact of the value of the statistical power

Suppose again that study #1 is designed with an *a priori* probability that the null-hypothesis is false of 20% ($\pi=0.20$). With a statistical power of 80% (by recruiting 79 dogs per group), the calculated FPRP is 20%. Suppose now that the investigators were finally able to recruit 39 dogs only per group instead of 79, the statistical power decreases to 50%. In this new situation, the calculated FPRP increases to 29% (see Figure 1). More generally, the lower the statistical power, the higher the FPRP. Since the statistical power of a study is directly related to its sample size, the FPRP increases with decreased sample size. This point indicates that a low statistical power (or small sample size) not only decreases the chances of obtaining a significant result

when there is a true association, but it also makes any obtained significant result more likely to be false positive.

C. Impact of the *a priori* probability that the null-hypothesis is false

Suppose again that study #1 is designed with a statistical power of 80% (with 79 dogs per group) and that the *a priori* probability that the null-hypothesis is false is 20% (π=0.20). The calculated FPRP is 20%. Suppose that study #2 is designed with 80% of statistical power as well (with 40 dogs per group), but with the *a priori* probability that the null-hypothesis is false of 1% (π=0.01), a low value due to its exploratory status. With such characteristics, the calculated FPRP of study #2 is 86% (see Figure 1). More generally, the lower the *a priori* probability that the null-hypothesis is false, the higher the FPRP. This point indicates that an exploratory study obtaining a significant association in the study sample is more likely to wrongly conclude that the association truly exists compared to a confirmatory study. This reasoning above is well known by clinicians using diagnostic tests[45,46]: a diagnostic test can have excellent sensitivity and specificity but have a very low PPV (and therefore a high FPRP) if the disease prevalence rate is very low. With statistical tests, a similar phenomenon occurs: a statistical test can be very sensitive (excellent statistical power) and very specific (low threshold of type-I error α), however a significant association in the study sample can very poorly predict the existence of a true association in the population if the *a priori* probability that this true association exists is very low.

The most difficult task to appreciate the probability of wrongly concluding that the association truly exists in the population (i.e., the value of the FPRP) is to appreciate the *a priori* probability that the null-hypothesis is false[32]. Such appreciation is beyond the scope of the paper. Briefly, some authors suggested a "reverse-Bayes" reasoning[47], which consists in setting the statistical power of the planned study and the desired FPRP value, then in seeing whether the value of the

a priori probability that the null-hypothesis is false is compatible with the current state of knowledge in the field[36,48].

X. CLINICAL SUMMARY

The probability of wrongly concluding that there is a true association in the population between one exposure and one outcome after obtaining a significant association in the study sample is neither equal to α nor to the p-value. Such probability (the "false positive report probability", or FPRP) depends on the characteristics of the study, namely its statistical power and its "confirmatory" *versus* "exploratory" status based on previous results in the same field.

In the case of an exploratory study (i.e., no previous studies have studied the same association in the same population yet), which is not uncommon in veterinary clinical research, researchers cannot be convinced that there is a true association in the population after obtaining a significant one in their study sample. However, in a study designed to confirm a result that other high-quality studies have previously obtained, researchers can start to be confident when they conclude that there is a true association after obtaining a significant one in their study sample.

Furthermore, and importantly, unless the true association is strong, a small sample size (which is also not uncommon in veterinary clinical research) prevents one from being confident when concluding, even after obtaining a significant association in the study sample.

Researchers can start to be confident when they conclude that there is a true association in the population between one exposure and one outcome after obtaining a significant one in the study sample if (1) the statistical power of the study is high (at least 80%), and (2) the amount of knowledge on the subject of the study allows to estimate the *a priori* probability that this association truly exists of at least 20%. In such a situation, when the association is significant, the probability of wrongly concluding that this association truly exists is 20% (FPRP=20%).

One may think that such probability of 20% is too elevated, compared to the 5% that most researchers have in mind when they conclude after obtaining a significant association. However, to reach an FPRP of 5%, the study must have the following characteristics: having a statistical power of 80% and being confirmatory with an *a priori* probability that there is a true association as high as 54% (π=0.54). Unfortunately, such confirmatory studies would not likely be designed because researchers would be concerned that most funding sources and journals would prioritize projects as being more innovative[3,49]. This is the reason why confirmatory studies should be much more encouraged than they actually are[50].

The readers must keep in mind that the interpretation of the probability of wrongly concluding that there is a true association in the population after obtaining a significant one in the study sample (i.e., the FPRP) assumes the absence of p-hacking, HARKing, association biases, or any other errors in scientific reporting. Even in this ideal world, the calculation of the FPRP, as it has been presented in this paper (and in other ones[10,35]), is still likely to be too optimistic[48] (i.e., the FPRP is likely to be even higher). Researchers as well as clinicians must nonetheless be aware that a significant association often provide a weak evidence that this association truly exists in the population. Such awareness is a necessary step to communicate more cautiously when writing the clinical relevance of the results of one study, potentially leading to clinical decisions thereafter. More generally, it is a necessary step for better veterinary research and (self-)evaluation of scientific research when practicing evidence-based veterinary medicine.

Acknowledgments

I wish to thank Prof. Fanny Storck and Dr. Elodie Darnis for their very helpful comments they provided, and Dr. Jeremy Beguin for helping me in finding illustrative examples in the field of oncology. I am finally grateful for the very helpful comments and suggestions by the editor and two referees on the originally submitted version.

References

1. Peng R. The reproducibility crisis in science: A statistical counterattack. *Significance* 2015;12:30-32.

2. Barba LA. The hard road to reproducibility. *Science* 2016;354:142.

3. Munafo MR, Nosek BA, Bishop DVM, et al. A manifesto for reproducible science. *Nat Hum Behav* 2017;1:0021.

4. Baker M. 1,500 scientists lift the lid on reproducibility. *Nature* 2016;533:452-454.

5. Collins FS, Tabak LA. Policy: NIH plans to enhance reproducibility. *Nature* 2014;505:612-613.

6. Begley CG. Six red flags for suspect work. *Nature* 2013;497:433-434.

7. Perrin S. Preclinical research: Make mouse studies work. *Nature* 2014;507:423-425.

8. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. *Nature* 2012;483:531-533.

9. Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. *Circ Res* 2015;116:116-126.

10. Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;2:e124.

11. Nuzzo R. Scientific method: statistical errors. *Nature* 2014;506:150-152.

12. Greenland S. Multiple comparisons and association selection in general epidemiology. *Int J Epidemiol* 2008;37:430-434.

13. Guller U, DeLong ER. Interpreting statistics in medical literature: a vade mecum for surgeons. *J Am Coll Surg* 2004;198:441-458.

14. Head ML, Holman L, Lanfear R, et al. The extent and consequences of p-hacking in science. *PLoS Biol* 2015;13:e1002106.

15. Bender R, Lange S. Adjusting for multiple testing--when and how? *J Clin Epidemiol* 2001;54:343-349.

16. Kerr NL. HARKing: hypothesizing after the results are known. *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc* 1998;2:196-217.

17. Delgado-Rodriguez M, Llorca J. Bias. *J Epidemiol Community Health* 2004;58:635-641.

18. Lanyon L. Evidence-based veterinary medicine: a clear and present challenge. *Vet Rec* 2014;174:173-175.

19. Vandeweerd JM, Kirschvink N, Clegg P, et al. Is evidence-based medicine so evident in veterinary research and practice? History, obstacles and perspectives. *Vet J* 2012;191:28-34.

20. White BJ, Larson RL. Systematic evaluation of scientific research for clinical relevance and control of bias to improve clinical decision making. *Journal of the American Veterinary Medical Association* 2015;247:496-500.

21. Kelsey JL. A contrary view on statistical significance. *Journal of the American Veterinary Medical Association* 2011;239:428-429.

22. West CP, Dupras DM. 5 ways statistics can fool you-Tips for practicing clinicians. *Vaccine* 2013;31:1550-1552.

23. Mullin CM, Arkans MA, Sammarco CD, et al. Doxorubicin chemotherapy for presumptive cardiac hemangiosarcoma in dogs. *Veterinary and comparative oncology* 2016;14:e171-e183.

24. Holtermann N, Kiupel M, Kessler M, et al. Masitinib monotherapy in canine epitheliotropic lymphoma. *Veterinary and comparative oncology* 2016;14 Suppl 1:127-135.

25. Shott S. Detecting statistical errors in veterinary research. *Journal of the American Veterinary Medical Association* 2011;238:305-308.

26. Lehmann EL. The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? *J Am Stat Assoc* 1993;88:1242-1249.

27. Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? *Bmj* 2001;322:226-231.

28. Jeffery N. Liberating the (data) population from subjugation to the 5% (P-value). *J Small Anim Pract* 2015;56:483-484.

29. McShane B, Gal D, Gelman A, et al. Abandon Statistical Significance. *Am Stat* 2019;73:235-245.

30. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* 2019;567:305-307.

31. White BJ, Larson RL, Theurer ME. Interpreting statistics from published research to answer clinical and management questions. *J Anim Sci* 2016;94:4959-4971.

32. Browner WS, Newman TB. Are all significant P values created equal? The analogy between diagnostic tests and clinical research. *JAMA* 1987;257:2459-2463.

33. Greenland S. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *Int J Epidemiol* 2006;35:765-775.

34. Lash TL. The Harm Done to Reproducibility by the Culture of Null Hypothesis Significance Testing. *Am J Epidemiol* 2017;186:627-635.

35. Wacholder S, Chanock S, Garcia-Closas M, et al. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst* 2004;96:434-442.

36. Held L. Reverse-Bayes analysis of two common misinterpretations of significance tests. *Clinical trials (London, England)* 2013;10:236-242.

37. Goodman SN. p values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am J Epidemiol* 1993;137:485-496; discussion 497-501.

38. Gliner JA, Leech NL, Morgan GA. Problems With Null Hypothesis Significance Testing (NHST): What Do the Textbooks Say? *J Exp Educ* 2002;71:83-92.

39. Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol* 2016;31:337-350.

40. Goodman S. A dirty dozen: twelve p-value misconceptions. *Seminars in hematology* 2008;45:135-140.

41. Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. *Am Stat* 2016;70:129-133.

42. Wagenmakers EJ. A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review* 2007;14:779-804.

43. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nat Hum Behav* 2018;2:6-10.

44. Trafimow D, Amrhein V, Areshenkoff CN, et al. Manipulating the Alpha Level Cannot Cure Significance Testing. *Frontiers in psychology* 2018;9:699.

45. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. *BMJ* 1994;309:102.

46. Grimes DA, Schulz KF. Uses and abuses of screening tests. *Lancet* 2002;359:881-884.

47. Matthews RAJ. Why should clinicians care about Bayesian methods? *J Stat Plan Inference* 2001;94:43-58.

48. Colquhoun D. The reproducibility of research and the misinterpretation of p-values. *Royal Society open science* 2017;4:171085.

49. Ten Hagen KG. Novel or reproducible: That is the question. *Glycobiology* 2016;26:429.

50. Mogil JS, Macleod MR. No publication without confirmation. *Nature* 2017;542:409-411.

Figure 1. Values of the false positive report probability (FPRP) according to the statistical power of a study (1-β) and the a priori probability that the null-hypothesis is false (π), for an type-I error set at 5% (α=0.05).

	A priori probability that H0 is false (π)								
Statistical Power (1-β)	1%	5%	10%	20%	30%	40%	50%	60%	70%
10%	98%	90%	82%	67%	54%	43%	33%	25%	18%
20%	96%	83%	69%	50%	37%	27%	20%	14%	10%
30%	94%	76%	60%	40%	28%	20%	14%	10%	7%
40%	93%	70%	53%	33%	23%	16%	11%	8%	5%
50%	91%	66%	47%	29%	19%	13%	9%	6%	4%
60%	89%	61%	43%	25%	16%	11%	8%	5%	3%
70%	88%	58%	39%	22%	14%	10%	7%	5%	3%
80%	86%	54%	36%	20%	13%	9%	6%	4%	3%
90%	85%	51%	33%	18%	11%	8%	5%	4%	2%

Appendix

To illustrate Formula (2), suppose the data shown in Table 4 from a hypothetical study of 1493 animals, including 20 diseased animals and 66 animals with a positive test result for a diagnostic test.

By using the classical formulas of Se, Sp, and PPV (see formulas under Table 1), we obtain:

$$Se = \frac{TP}{TP + FN} = \frac{12}{15} = 0.80$$

$$Sp = \frac{TN}{FP + TN} = \frac{1404}{1478} = 0.95$$

$$PPV = \frac{TP}{FP + TP} = \frac{12}{86} = 0.14$$

In order to calculate the PPV of the diagnostic test from Formula (2), we need to calculate π, the proportion of diseased animals: $\pi = 15/1493 = 0.01$. By including the values of Se, Sp, and π in Formula (2), we obtain the same value for the PPV of the diagnostic test as the previously calculated one:

$$PPV = \frac{0.01 \times 0.80}{0.01 \times (1 - 0.95) + 0.01 \times 0.80} = 0.14$$

Table 1. Distribution of frequencies within a sample size of N animals according to the results of a diagnostic test and the absence/presence of the disease.

Result of a diagnostic test	Disease present	Disease absent	Total
Positive	True Positive (TP)	False Positive (FP)	FP+TP
Negative	False Negative (FN)	True Negative (TN)	TN+FN
Total	TP+FN	FP+TN	N

Table 2. Distribution of frequencies within a sample size of N animals according to sensitivity (Se) and specificity (Sp) of a diagnostic test, and according to the proportion π of diseased animals in the sample.

Result of a diagnostic test	Disease present	Disease absent	Total
Positive	$N.\pi.Se$	$N.(1-\pi).(1-Sp)$	$N.((1-\pi).(1-Sp)+\pi.Se)$
Negative	$N.\pi.(1-Se)$	$N.(1-\pi).Sp$	$N.((1-\pi).Sp+\pi.(1-Se))$
Total	$N.\pi$	$N.(1-\pi)$	N

Table 3. Analogies between diagnostic and statistical tests.

Situation of diagnostic tests	Corresponding situation for statistical tests
"To have the disease"	"There is a true association in the population" ("the null-hypothesis is false")
"To be disease-free"	"There is no true association in the population" ("the null-hypothesis is true")
"To obtain a positive result"	"To obtain a significant association" ("$p \leq \alpha$")
"To obtain a negative result"	"To obtain a non-significant association" ("$p > \alpha$")

Table 4. Illustrative example of distribution of frequencies within a sample size of 1493 animals.

Result of a diagnostic test	Disease present	Disease absent	Total
Positive	12	74	86
Negative	3	1404	1407
Total	15	1478	1493