Challenges of making decisions on the basis of significant statistical associations

Loic Desquilbet PhD

From the Biostatistics and Clinical Epidemiology Service, Ecole Nationale Vétérinaire d'Alfort, and U955 Institut Mondor de Recherche Biomédicale, Institut National de la Santé et de la Recherche Médicale, Université Paris Est Créteil, Maisons-Alfort, F-94700, France.

Address correspondence to Dr. Desquilbet (loic.desquilbet@vet-alfort.fr).

A crisis has recently been described in the reproducibility of studies reported in leading science journals.[1–6] One of the multiple origins of this crisis is that significant statistical results obtained in some studies were not replicated or supported in other studies.[7–9] One of the reasons given is that researchers often wrongly generalize their results to the population of interest (ie, the target population) after obtaining a significant result in their study sample.[10,11] *P*-hacking and HARKing (where HARK stands for "hypothesis after result is known") are inappropriate methods of analyzing and interpreting study findings that lead to false-positive results. *P*-hacking (also called *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 test approaches until nonsignificant 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, biases,[17] or any other errors in scientific reporting, the probability of wrongly inferring that obtained significant results apply to the target population is high in some situations.

The purpose of this article is to help researchers in clinical veterinary science and clinicians who interpret research findings appreciate the degree of certainty or uncertainty when generalizing the study results to the target population of studied animals, according to the characteristics of the clinical study. This appreciation is also important in the practice of evidence-based veterinary medicine,[18] particularly when critically appraising the evidence within the more general framework of the clinical decision-making process.[19,20] To this end, 2 examples of hypothetical studies (each considered feasible and ethical for simplification) will be used. Statistical and diagnostic test concepts will be reviewed to demonstrate the similarity between diagnostic test interpretation and statistical test interpretation and to illustrate the calculation of the probability of an incorrect conclusion after a significant association has been obtained.

General Concepts

Throughout this article, the context will always be the following: researchers seek to provide evidence that there is a true association between an exposure (eg, neuter status, a certain treatment [vs placebo], or a surgical [vs medical] intervention) and an outcome (eg, disease occurrence, tumor remission rate, or survival rate). The statistical tests mentioned are those that can be used to test such associations, and the conclusions drawn from other statistical tests (eg, tests for normality of data distribution) will not be addressed. Furthermore, the term "significant association" will be used to describe an association classified as significant on the basis of the observed *P* value ($P < \alpha$) and not on the basis of clinical importance.[21,22]

Hypothetical Studies

The first hypothetical study (study 1) used as an example to explain concepts is as follows. Mullin et al[23] conducted a nonrandomized study to assess the association between doxorubicin chemotherapy (vs no chemotherapy; control group) and survival time following diagnosis in dogs with presumptive cardiac hemangiosarcoma. The observed significant difference in survival times between the 2 groups suggested a potential effect of doxorubicin chemotherapy on survival time. Building on this information, another group of investigators design a randomized placebo-controlled clinical trial (study 1) to confirm the beneficial effect of doxorubicin chemotherapy within the first 4 months of use. To do so, they use the Kaplan-Meier curves provided in the previous report,[23] which show that 45% of dogs in the doxorubicin group and 5% of dogs in the control group remained alive 4 months after diagnosis, and use these data to calculate a sample size sufficient to yield 80% statistical power (ie, 79 dogs/group). The investigators then follow the dogs for 4 months after placebo or treatment initiation and compare survival times during this period between groups with the Kaplan-Meier method and log-rank test. Study 1 can be considered confirmatory because the purpose is to confirm the result of the previous study.

In the second example (study 2), masitinib monotherapy has been suggested to have potential for the treatment of dogs with epitheliotropic T-cell lymphoma.[24] Consequently, investigators seek to conduct a randomized controlled clinical trial to assess the effect of masitinib in dogs with multicentric lymphoma on remission (partial or complete) rate. To do so, they randomly allocate 80 dogs with multicentric lymphoma to receive masitinib plus prednisone (n = 40; mastinib group) or prednisone only (40; control group). This number (n = 80) was not based on a priori sample size calculation, but rather on the number of dogs that the investigators anticipated they could enroll during the predefined period. They then follow the dogs for 3 months after treatment begins and compare between groups the proportion of dogs that achieve partial or complete remission by that time. Study 2 can be considered exploratory because it is the first to assess a potential association between masitinib plus prednisone use and lymphoma remission rate in the population of dogs with multicentric lymphoma.

Review of Statistical Concepts

The null hypothesis and its acceptance or rejection

A statistical test of the association between an exposure and outcome is based on a null hypothesis, which states that there is no association in a predefined (target) population.[25] For instance, the null hypothesis of the log-rank test performed in study 1 is that there is no association between doxorubicin chemotherapy (vs a placebo) and survival time in dogs with presumptive cardiac hemangiosarcoma. If the null hypothesis is rejected, the conclusion is that the study provides evidence in support of a true association in the population between the exposure and the outcome. Conventionally, rejection or acceptance of the null hypothesis has been based on the *P* value yielded by the statistical test. If that *P* value is less than a threshold value (α), the association is classified as significant in the study sample, and one concludes that the null hypothesis is false (ie, the null hypothesis is rejected). Nevertheless, this use of *P* values and the significant-nonsignificant approach has been questioned by many scientists[26–29] and is believed to have contributed to the reproducibility crisis.

Type I and type II errors and statistical power

When no true association exists between the exposure and outcome in the population (ie, when the null hypothesis is true), the probability of obtaining a significant ($P < \alpha$) association in the study sample is α (also known as the type I error rate). When a true association exists between the exposure and outcome in the population (ie, when the null hypothesis is false), the probability of not obtaining a significant association in the study sample is equal to β (also known as the type II error rate). Therefore, in such a situation, the probability of obtaining a significant association is $1 - \beta$, which represents the statistical power of the study.

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.[30] The positive predictive value (PPV) of a diagnostic test is the probability that an individual with a positive test result would truly have the disease.[30] In a sample of N animals, Se, Sp, and PPV can be estimated by calculating the proportion of true-positive animals (ie, those for which a positive test result is true) among diseased animals (Se), the proportion of true-negative animals (ie, those for which a negative test result is true) among disease-free animals (Sp), and the proportion of true-positive animals among test-positive animals (PPV) as follows:

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

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

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

where TP is the number of true-positive animals, FN is the number of false-negative animals, FP is the number of false-positive animals, and TN is the number of true-negative animals **(Table 1)**.

These values can be reexpressed by using proportions instead of frequencies. To do so, let π be the prevalence of the disease in the population, which represents the probability of having the disease for an animal randomly drawn from the population before the result of the diagnostic test is known. If a sample of N animals is randomly drawn from a population in which the prevalence of disease is π, it is expected that there will be $N \times \pi$ diseased animals and $N \times (1 - \pi)$ disease-free animals. Among the $N \times \pi$ diseased animals, there will be $N \times \pi \times$ Se true-positive animals; among the $N \times (1 - \pi)$ disease-free animals,

Table 1—Distribution of frequencies (ie, number of animals) within a sample of N animals according to the results of a diagnostic test and the true absence or presence of the disease.

Result	Disease present	Disease absent	Total
Positive	TP	FP	**FP + TP**
Negative	FN	TN	**TN + FN**
Total	**TP + FN**	**FP + TN**	**N**

FN = The number of false-negative animals. FP = The number of false-positive animals. TN = The number of true-negative animals. TP = The number of true-positive animals.

there will be N × (1 – π) × Sp true-negative animals **(Table 2)**.

The formula for calculation of PPV can be accordingly reexpressed by replacing TP with N × π × Se and FP with N × (1 – π) × (1 – Sp), which yields the following formula:

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

The equivalence between the 2 formulas for PPV (ie, formulas 1 and 2) can be demonstrated in a hypothetical example involving 1,493 animals, 15 of which truly have the disease of interest and 86 of which have a positive result from a diagnostic test, yielding an Se of 0.80 and Sp of 0.95 **(Table 3)**. The PPV as calculated with formulas 1 and 2, respectively, is as follows:

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

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

These calculations show that, for situations in which the prevalence of the disease is low (eg, 1%), a diagnostic test that has very good Se (eg, 80%) and excellent Sp (eg, 95%) will still have poor PPV (eg, 14% per both formulas for PPV). A PPV of 14% would be interpreted as indicating that only 14% of animals with a positive test result truly have the disease, whereas 86% (100% – 14%) of animals with a positive test result are truly disease free.

Application of Diagnostic Test Concepts to Statistical Tests

Although fundamental differences exist between interpretation of diagnostic test results and interpretation of statistical test results, some aspects of diagnostic test interpretation can serve as a reasonable analogy for statistical test interpretation.[31] To apply the concepts of Se, Sp, and PPV to a statistical test of the association between an exposure and outcome, one must define analogous terms for having the disease, being disease-free, and having positive and negative test results. With a statistical test, researchers typically seek evidence that there is a true association in the population (ie, evidence supporting that the null hypothesis is false), and a significant association in the study sample is in favor of the evidence they are seeking. In the situation of diagnostic tests, clinicians may seek evidence that an animal has a certain disease, and a positive test result for this disease is in favor of the evidence they are seeking **(Table 4)**.[32] The Se of a diagnostic test is the probability of obtaining a positive result when the animal has the disease. By analogy, the Se of a statistical test is therefore the probability of obtaining a significant association when the null hypothesis is false, which is analogous to the statistical power of a study (1 – β). The Sp of a diagnostic test is the probability of obtaining a negative result when the animal is disease free. By analogy, the specificity of a statistical test is the probability of obtaining a nonsignificant association when the null hypothesis is true. Given that α is the probability of obtaining a significant association when the null hypothesis is true, by analogy, the Sp of a statistical test is equivalent to 1 – α.

Relevance of calculating the PPV of a statistical test

In veterinary practice, the PPV of diagnostic tests is much more useful to clinicians than are the Se and Sp of such tests. When a positive test result is obtained for an animal, clinicians would typically like to know the probability that the animal truly has the disease, which the PPV reflects. By analogy with statistical tests, when a significant association is obtained, one would like to know the probability that the association truly exists in the population (ie, the PPV of 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 and survival time for dogs with presumptive cardiac hemangiosarcoma, given that the investigators obtained a significant association in their study sample. When the PPV of a statistical test is low, the probability that there is a true association in the population is low even when the association was deemed by the test to be significant. In such situations of low PPV, it would not be

Table 3—Illustrative example of distribution frequencies of hypothetical diagnostic test results for a sample size of 1,493 animals in which the prevalence of disease is 1% (ie, 15 animals truly have the disease).

Result	Disease present	Disease absent	Total
Positive	12	74	**86**
Negative	3	1,404	**1,407**
Total	**15**	**1,478**	**1,493**

$$Se = \frac{TP}{TP + FN} = \frac{12}{15} = 0.80 \qquad Sp = \frac{TN}{FP + TN} = \frac{1,404}{1,478} = 0.95$$

Table 2—Distribution of expected frequencies within a sample size of N animals randomly drawn from a population in which the prevalence of a disease is π, and according to Se and Sp of a diagnostic test.

Result	Disease present	Disease absent	Total
Positive	N × π × Se	N × (1 – π) × (1 –Sp)	**N × ([1 – π] × [1 – Sp] + [π × Se])**
Negative	N × π × (1 – Se)	N × (1 – π) × Sp	**N × ([{1 – π} × Sp] + [π × {1 – Se}])**
Total	**N × π**	**N × (1 – π)**	**N**

Table 4—Analogies between diagnostic tests and statistical tests.

Diagnostic test interpretation	Statistical test interpretation
This animal has the disease	There is a true association in the population (ie, the null hypothesis is false)
This animal is disease free	There is no true association in the population (ie, the null hypothesis is true)
Positive test result	Significant ($P < \alpha$) association
Negative test result	Nonsignificant ($P > \alpha$) association
Se, or probability of obtaining a positive test result when the animal is truly diseased	Probability of obtaining a significant association when the null hypothesis is false (ie, power or $1 - \beta$)
Sp, or probability of obtaining a negative test result when the animal is truly disease free	Probability of obtaining a nonsignificant association when the null hypothesis is true (ie, $1 - \alpha$)
PPV, or probability that an animal with a positive test result truly has the disease	Probability that the significant association obtained between the exposure and outcome in the study sample truly exists in the population

surprising if the significant association identified in one study could not be replicated in another study of the same association targeting the same population.[10]

To calculate the PPV of a statistical test, values for Se, Sp, and π in formula 2 can be replaced with their analogous values for statistical tests (Table 4). 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 a study to test the association between an exposure and outcome, they have some level of uncertainty that this association truly exists in the population. They must have such a level of uncertainty because a study would not be necessary if they knew with 100% certainty that the null hypothesis is false. Before conducting a study, researchers therefore have in mind an a priori probability that the null hypothesis is false in the studied population, and this probability lies between but excludes 0 and 1. For instance, because the association between doxorubicin chemotherapy and time to death in dogs with presumptive cardiac hemangiosarcoma had been previously suggested, the investigators of study 1 should have a higher level of certainty that this association truly exists, compared with that of the investigators of study 2, where the tested association had never been evaluated 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. For a statistical test, π would be the a priori (ie, before the statistical test is performed) probability that the null hypothesis is false (ie, the a priori probability that the association truly exists in the population). The expression "a priori" 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[33] wrote, the value of π for statistical tests is based on "biologic plausibility, previous experience with similar hypotheses, and knowledge of alternative scientific explanations." Therefore, in an exploratory study where researchers are the first ones to assess an association between an exposure and 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 pathophysiologic basis for this exploratory study.[10]

Formula 2 for the PPV of a diagnostic test can be rewritten by replacing Se and Sp with their analogous values for statistical tests (ie, [1 – α] and [1 – β], respectively) as follows:

$$\text{PPV} = \frac{\pi \times (1 - \beta)}{([1 - \pi] \times \alpha) + (\pi \times [1 - \beta])} \quad \text{(formula 3)}$$

where π represents the a priori probability that the null hypothesis is false (ie, the probability that the association truly exists), α is the type I error rate, β is the type II error rate, and (1 – β) is the statistical power.

For example, suppose that the statistical power of study 2 is 80% when testing the association between masitinib plus prednisone use (vs prednisone use only) and lymphoma remission rate in dogs with multicentric lymphoma and that α is set at 5% (ie, α = 0.05). Suppose that the probability that this association truly exists is 1% (ie, 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 an association. On the basis of these characteristics, the PPV as calculated with 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%.

False-positive report probability of a statistical test

For situations in which researchers would like to estimate the probability of wrongly concluding that there is a true association in the population after obtaining a significant association in the study sample, the complement of the PPV (1 – PPV) is the most relevant indicator. This complement of PPV is called the false-positive report probability (FPRP)[35,36] and is 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 association 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 and survival time for

dogs with presumptive cardiac hemangiosarcoma after obtaining a significant association in the study sample.

The formula for FPRP can be derived from formula 3 as follows:

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

Numeric examples of FPRP values according to selected values of π and (1 – β) are provided **(Figure 1)**.

Misinterpretation of the Type I Error Rate and *P* Value

The type I error rate in statistical testing is typically set at 5% (α = 0.05), which is considered a low value. Many researchers wrongly believe that because α is low, the conclusion regarding a significant association is accompanied by a corresponding low probability of error.[37,38] Similarly, *P* values are commonly misinterpreted as the observed probability of wrongly rejecting 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 erroneously that there is strong evidence for a true association between doxorubicin chemotherapy and survival time of dogs with presumptive cardiac hemangiosarcoma, with a 3% risk of error.[27,40,41] However, the *P* value is actually the probability of the observed or more extreme results if the null hypothesis is true and if no bias existed when estimating the association. Therefore, the *P* value has no meaningful interpretation per se because its value is conditional on a hypothesis that nobody knows with 100% certainty is true or false.[42]

Factors Contributing to a High FPRP

Formula 4 indicates that the FPRP of a statistical test performed in a study designed to provide evidence that there is a true association between an exposure and outcome depends on the type I error rate (α), statistical power of the study (1 – β), and a priori probability that the null hypothesis is false (π).

Impact of the value for type I error rate

Suppose that study 1 is designed to have 80% statistical power, and the a priori probability that the null hypothesis is false is 20% (ie, π = 0.20). If α is set to 1%, the FPRP calculated from formula 4 is 5%; if α is set to 5%, the FPRP increases to 20% (Figure 1). Therefore, and more generally, the higher the type I error rate is set, the higher the FPRP will be. This point is one of the origins of a scientific movement that questions the type I error threshold of 5% and proposes to lower it to 0.5% (α = 0.005).[43] Nevertheless, this thinking is not shared by all scientists,[44] and the convention for setting the type I error threshold at 5% is likely to persist for years. Consequently, an α value of 0.05 is used in all examples that follow.

Impact of the value for statistical power

Suppose again that study 1 is designed with an a priori probability of 20% that the null hypothesis is false. With a statistical power of 80% (by recruiting 79 dogs/group), the calculated FPRP is 20%. Suppose now that the investigators are able to recruit only 39 dogs/group, reducing the statistical power to 50%. In this new situation, the calculated FPRP increases to 29% (Figure 1). More generally, the lower the statistical power is, the higher the FPRP will be. Because the statistical power of a study is directly related to its sample size, the FPRP increases with decreasing sample size, indicating that a low statistical power (or small sample size) not only decreases the chance of obtaining a significant result when there is a true association, but it also makes any obtained significant result more likely to be falsely positive.

Statistical power (1 – β)	A priori probability that H0 is false (π)								
	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%

Figure 1—Values of FPRP according to the statistical power (1 – β) of a study and the a priori probability that the null hypothesis (H0) is false (π) when the type I error rate is set at 5% (α = 0.05). The darker the shading, the higher the FPRP (ie, the higher the probability of wrongly concluding that there is a true association between an exposure and outcome in the population after obtaining a significant association in the study sample).

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/group), with an a priori probability of 20% that the null hypothesis is false and a calculated FPRP of 20%. Suppose that study 2 is also designed with 80% statistical power (with 40 dogs/group), but with an a priori probability of 1% that the null hypothesis is false, which is low owing to the exploratory nature of that study. With such characteristics, the calculated FPRP of study 2 is 86% (Figure 1). More generally, the lower the a priori probability is that the null hypothesis is false, the

higher the FPRP will be, which indicates that an exploratory study with a significant association obtained in the study sample is more likely to wrongly lead to the conclusion that the association truly exists than is a confirmatory study. Indeed, this reasoning applies to diagnostic tests[30,45]: a diagnostic test can have excellent Se and Sp but a very low PPV (and therefore a high FPRP) if the disease prevalence is very low. Similarly, a statistical test can be very sensitive (excellent statistical power) and very specific (low type I error rate); however, a significant association in the study sample would very poorly predict the existence of a true association in the population if the a priori probability that a true association exists is very low.

The most difficult component to determine when estimating the probability of wrongly concluding that the association truly exists in the population (ie, the value of the FPRP) is the a priori probability that the null hypothesis is false.[32] An explanation of how this probability might be derived is beyond the scope of this article. Briefly, some authors suggest the use of reverse-Bayes reasoning,[46] which consists of setting the statistical power of the planned study and the desired FPRP value, then determining 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.[36,47]

Clinical Summary

The probability of wrongly concluding that there is a true association between an exposure and outcome in the population after obtaining a significant association in the study sample is equal to neither α nor the *P* value. Such probability (the FPRP) depends on the characteristics of the study, namely its statistical power and whether its purpose is confirmatory or exploratory, given the existing research findings on the subject.

In exploratory studies (ie, those involving associations that have not previously been evaluated in the same population), which are not uncommon in veterinary clinical research, researchers cannot and should not be convinced that there is a true association between an exposure and outcome in the population after obtaining a significant association in their study sample. However, in a study designed to confirm a significant result obtained in previous high-quality studies, researchers can have more confidence in this regard. Furthermore and importantly, unless the true association is strong, a small sample size (which is also not uncommon in veterinary clinical research) necessarily prevents researchers from being confident in their conclusions about their study sample, even after a significant association has been obtained.

Researchers can start to have confidence that there is a true association between an exposure and outcome in the population after obtaining a significant association in the study sample if 1) the statistical power of the study is high (at least 80%) and 2) the existing information on the subject indicates that the a priori probability that this association truly exists is at least 20%. In such a situation, when the tested association is significant, the probability of wrongly concluding that it truly exists (FPRP) is 20%. One may believe that a probability of 20% is too high, compared with the 5% that most researchers have in mind when making conclusions about a significant association. However, to achieve an FPRP of 5% when a study has a statistical power of 80% would require that the study be confirmatory, with an a priori probability of a true association of 54% (Figure 1). Unfortunately, such confirmatory studies would not likely be designed because researchers would then be concerned that most funding sources and journals would prioritize other, more novel projects.[3,48] This is the reason that confirmatory studies should be encouraged more than they actually are.[49]

Those interpreting research findings 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 assumes the absence of *P*-hacking, HARKing, biases, or any other errors in scientific reporting. Even in the ideal situation, the FPRP, as calculated in this article and others,[10,35] is still likely to be too optimistic[47] and the true FPRP is likely to be even higher. Researchers and clinicians must nonetheless be aware that although an association might be identified as significant in a study, this is often weak evidence that the association truly exists in the population. Such awareness is a necessary step toward more cautious communication about the clinical relevance of the results of a study, potentially leading to clinical decisions thereafter. More generally, it is also a necessary step toward better veterinary research and evaluation of scientific research when practicing evidence-based veterinary medicine.

Acknowledgments

The author thanks Professor Fanny Storck and Dr. Elodie Darnis for their helpful comments and Dr. Jeremy Beguin for assistance in finding illustrative examples in the field of oncology.

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. *Pers Soc Psychol Rev* 1998;2:196–217.
17. Delgado-Rodríguez 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. *J Am Vet Med Assoc* 2015;247:496–500.
21. Kelsey JL. A contrary view on statistical significance. *J Am Vet Med Assoc* 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. *Vet Comp Oncol* 2016;14:e171–e183.
24. Holtermann N, Kiupel M, Kessler M, et al. Masitinib monotherapy in canine epitheliotropic lymphoma. V*et Comp Oncol* 2016;14(suppl 1):127–135.
25. Lehmann EL. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *J Am Stat Assoc* 1993;88:1242–1249.
26. Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? *BMJ* 2001;322:226–231.
27. Jeffery N. Liberating the (data) population from subjugation to the 5% (P-value). *J Small Anim Pract* 2015;56:483–484.
28. McShane B, Gal D, Gelman A, et al. Abandon statistical significance. *Am Stat* 2019;73:235–245.
29. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature* 2019;567:305–307.
30. Grimes DA, Schulz KF. Uses and abuses of screening tests. *Lancet* 2002;359:881–884.
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 1*987;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. *Clin Trials* 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. *Semin Hematol* 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. *Psychon Bull Rev* 2007;14:779–804.
43. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. *Nat Hum Beha*v 2018;2:6–10.
44. Trafimow D, Amrhein V, Areshenkoff CN, et al. Manipulating the alpha level cannot cure significance testing. *Front Psychol* 2018;9:699.
45. Altman DG, Bland JM. Diagnostic tests 2: predictive values. *BMJ* 1994;309:102.
46. Matthews RAJ. Why should clinicians care about Bayesian methods? *J Stat Plan Inference* 2001;94:43–58.
47. Colquhoun D. The reproducibility of research and the misinterpretation of p-values (Erratum published in *R Soc Open Sci* 2018;5:180100). *R Soc Open Sci* 2017;4:171085.
48. Ten Hagen KG. Novel or reproducible: that is the question. *Glycobiology* 2016;26:429.
49. Mogil JS, Macleod MR. No publication without confirmation. *Nature* 2017;542:409–411.