Power and Error: Increased Risk of False Positive Results in Underpowered Studies

R.M. Christley*

Epidemiology and Public Health, School of Veterinary Science, Faculty of Health and Life Science, University of Liverpool, Leahurst Campus, Neston, CH64 7TE, UK

Abstract: It is well recognised that low statistical power increases the probability of type II error, that is it reduces the probability of detecting a difference between groups, where a difference exists. Paradoxically, low statistical power also increases the likelihood that a statistically significant finding is actually falsely positive (for a given p-value). Hence, ethical concerns regarding studies with low statistical power should include the increased risk of type I error in such studies reporting statistically significant effects. This paper illustrates the effect of low statistical power by comparing hypothesis testing with diagnostic test evaluation using concepts familiar to clinicians, such as positive and negative predicative values. We also note that, where there is a high probability that the null hypothesis is true, statistically significant findings are even more likely to be falsely positive.

INTRODUCTION

Significance tests cannot determine whether a null hypothesis is true or not, they can only indicate the probability of observing the data collected assuming the null hypothesis is true. When using significance tests to make decisions about a null hypothesis, two types of error can be made; rejecting the null hypothesis when it is true (type I error) and failure to reject the null hypothesis when it is, in fact, false (type II error). It is well recognised that low statistical power increases the probability of a type II error. In contrast, if two experiments are conducted and each of the null hypotheses are rejected (with the statistical tests used having the same p-value), it is sometimes assumed that the type I error rates are the same in each case. However, two factors may influence interpretation of such situations. First, if one of the null hypotheses was highly likely to be true, the result of this experiment should be treated with greater caution. Also, if one of the studies had substantially lower statistical power, this study has an increased probability of incorrectly rejecting the null hypothesis. Here we illustrate how interpretation of hypothesis tests may vary with statistical power and the probability of the null hypothesis being true through comparison with methods of diagnostic test evaluation.

BACKGROUND

Discussion of the statistical power of medical studies and the potential effects of underpowered studies is widespread in the medical literature. Many published randomised trials [1-3] and observational studies [4] have low statistical power, or fail to calculate or report power analyses [3]. For example, a review of 127 randomised control trials in the surgical literature found that only half had sufficient power to detect large differences between treatment groups [5]. Another review of negative clinical trials and experimental studies in the plastic surgery literature reported that 98% of studies with binary responses had inadequate power to detect a 25% change in effect, whereas 85% of studies with continuous outcomes had inadequate power to detect a mean difference of one standard deviation [6]. Startlingly, a review of 117 randomised trials investigating fracture treatment estimated the average power of studies to be 25% with a range of 2 to 99% [7].

It is well recognised that low power increases the probability of type II error and this effect is sometimes cited by authors as a possible reason for no statistically significant result being identified. Underpowered studies have been labelled "scientifically useless", principally because low statistical power increases the risk of type II errors (failing to observe a difference when the null hypothesis is actually false) [3, 8-9]. Indeed, numerous authors have suggested that it may be unethical to involve people in epidemiological studies or clinical trials that have low statistical power because such studies may be inadequately able to test the hypotheses of interest [8-10].

However, there is second reason why studies with low statistical power are problematic and may be considered unethical: low statistical power leads to increased risk that statistically significant results will actually be falsely positive (for any given p-value). It is often assumed that the probability of a false positive significant result is given by the critical probability α, usually set at 0.05. However, for many studies the false positive error rate will be considerably greater than the predefined critical probability [11]. This effect is greatest when the study power and/or the probability of the null hypothesis being false are low [12]. That is, there is an increased risk that a statistically significant, but surprising (i.e. improbable), result is actually "falsely positive", and this effect is greatest in small studies. We illustrate this by comparing hypothesis testing with diagnostic test evaluation using concepts familiar to clinicians, such as positive and negative predicative values (and likelihood

*Address correspondence to this author at the Epidemiology and Public Health, School of Veterinary Science, Faculty of Health and Life Science, University of Liverpool, Leahurst Campus, Neston, CH64 7TE, UK;
Te: +44 (0) 151 794 6170; E-mail: robc@liverpool.ac.uk

1874-2971/10

ratios). Through comparison with diagnostic test evaluation we also emphasize the role that prior information or belief can play in our interpretation of hypothesis tests.

HYPOTHESIS TESTS AND DIAGNOSTIC TEST EVALUATION

Hypothesis testing (HT), proposed by Neyman and Pearson [13] almost 80 years ago, provides a framework for the interpretation of experimental results (Table **1**). This approach leads to the potential for two types of error; declaring that there is a difference between the groups, when in fact there is no (clinically important) difference (rejecting the null hypothesis when it is in fact true; type I or α error), and deciding there is no difference, when in fact a clinically important difference exists (type II or β error, which is equivalent to 1-power, where *power* is probability that a null hypothesis will be rejected when it is indeed false).

Table 1. Possible Errors that May Arise when Testing Experimental Hypotheses

Experimental Results	Truth	
	Difference Exists	No Difference
Reject H_0	Power = 1 - β	α
Do not reject H_0	β	-

β = false negative rate or the probability of a type II error.
α = false positive rate or the probability of a type I error.

There are similarities between this representation of HT with that of diagnostic test evaluation (Table **2**) [14] and revising the common representation of HT illustrated in Table **1** to that commonly used for diagnostic tests further highlights the similarity in logic between these methods (Table **3**). Indeed, the common representation of HT in Table **1** may be misleading, as there is no indication of the denominator values used in the calculation of the type I and type II error rates. In contrast, from Table **3** it is clear that the type I error rate equals b/(b+d), or FP/(FP+TN), whereas the type II error rate equals c/(a+c), or FN/(TP+FN).

When considering diagnostic tests, numerous parameters familiar to clinicians can be calculated (Fig. **1**). Several parameters, including the sensitivity, specificity and likelihood ratios for positive and negative test results, are intrinsic to the test, whereas the predictive values of the test results vary with disease prevalence. From Tables **2** and **3**, it is evident that the probability of a type I error is b/(b+d), equivalent to 1-specificity of the diagnostic test. Similarly, the power of the study (i.e. 1-β) is given by a/(a+c), which is equivalent to the sensitivity of the diagnostic test.

Table 2. The Possible Outcomes of Diagnostic Tests Compared to the True Disease Status

Diagnostic Test Results	Actual Condition		Total
	Disease Present	No Disease	
Positive	a	b	a+b
Negative	c	d	c+d
Total	a+c	b+d	N

a = true positive (TP); b = false positive (FP); c = false negative (FN); d = true negative (TN).

Table 3. The Possible Errors Outcomes that May Arise when Testing Experimental Hypotheses, Represented in the Same Format as Traditionally Used for Diagnostic Tests

Experimental Results	Actual condition		Total
	Difference Exists	No Difference	
Reject H_0	a	b	A+b
Do not reject H_0	c	d	C+d
Total	a+c	b+d	N

a = true positive (TP); b = false positive (FP) = type I error; c = false negative (FN) = type II error; d = true negative (TN).

Importantly, when interpreting a diagnostic test result, knowledge of the sensitivity and specificity of a test provide only limited information. Of greater interest is the probability that a particular test result is true; the predictive values. The predictive value of a positive test is the probability that an individual that tested positive actually has the disease in question. Similarly, the predicative value of a negative test is the probability that someone with a negative test result is actually free of the disease. The predictive values are related to the test sensitivity and specificity through the disease prevalence.

Sensitivity (Se):	a/(a+c)	= 1-β
Specificity (Sp):	d/(b+d)	= 1- α
Positive predictive value (PPV):	a/(a+b)	
Negative predictive value (NPV):	d/(c+d)	
Likelihood ratios for positive test (LR+):	(Se)/(1-Sp)	= (1- β)/ α
Likelihood ratios for negative test (LR-):	(1-Se)/(Sp)	= β /(1- α)

Fig. (1). The relationship between parameters associated with diagnostic tests and those associated with statistical testing. β = false negative rate or the probability of a type II error α = false positive rate or the probability of a type I error.

Similar issues arise when the results of an experimental study are interpreted. Whilst the sensitivity and specificity of the experiment can be calculated prior to the study (and is analogous to sample size calculation), estimation of the predictive value is less intuitive. However, it is the predictive values that are most revealing; what is the probability that there truly is an effect, given that the null hypothesis has been rejected, or that there is no effect, given that the null hypothesis has not been rejected?

THE PROBABILITY OF THE NULL HYPOTHESIS

The characteristics of the diagnostic test (sensitivity and specificity) are related to the predictive values through the disease prevalence. What is the equivalent parameter to disease prevalence for the hypothesis test? Sterne and Davey Smith [15] consider this probability to be the prevalence of truly positive results amongst a large number of tests. Here they assume that the study in question is just one of a population of studies, where a certain proportion of studies are of null hypotheses which are false.

With regard to diagnostic tests, we can also consider the probability that a particular individual will be disease positive, prior to undertaking the test. Hence an individual with clinical signs of a disease and exposure to known risk factors may be considered more likely to have a disease than someone free of clinical signs and with no known relevant exposures. If a test were positive for each of these individuals, we would be more likely to believe the test result for the former, and consider that the latter test result may be more likely to be a false positive. In effect, we assume that these individuals come from two populations, the former with a higher prevalence of disease than the latter. In this way, the probability of disease before the test was performed (i.e. the prior probability) at the level of the individual is the prevalence of disease in a population of similar individuals.

Similarly, for hypothesis testing, we can use our knowledge of the hypothesis in question to estimate the prior probability that the null hypothesis is true (that there is truly no difference). Assuming we have formulated the hypotheses *a priori* based on existing evidence of an effect it is likely that we would conclude that the probability of the null hypothesis (i.e. of no effect) being true is low, and rejecting the null hypothesis will provide strong support for this conclusion. If, however, we have no good reason to form an opinion regarding the null hypothesis, such as when the finding was a "surprise" result amongst many measured variables, the fact that we reject the null hypothesis may not necessarily lead us to markedly change our conclusions about the variable in question.

FALSE POSITIVE RATE OF STATISTICAL TESTS

Therefore, the predictive values of a statistical test (i.e. the probability that a particular conclusion is correct) are influenced by the prior probability of the null hypothesis. Fig. (**2**) shows the probability of a false positive conclusion when our study data indicates we should reject the null hypothesis, assuming a study with type I error rate of 0.05 and type II error rate ranging from 0.2 to 0.8. It is clear that as the prior probability of a difference (i.e. H_0 is false) increases, the chance of a false positive decision decreases. It is worth noting that the commonly accepted values for α and

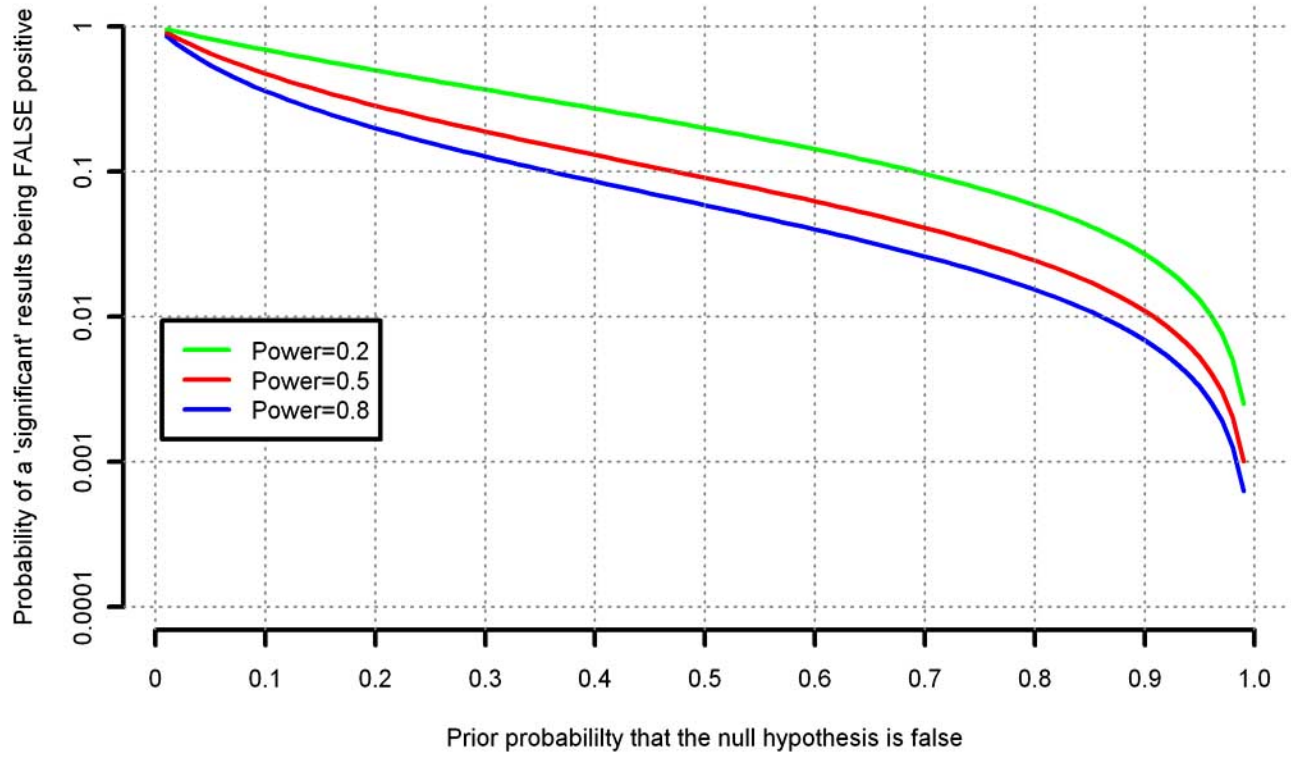

Fig. (2). The effect of varying the statistical power of a study on the probability of a false positive significant result (assuming alpha=0.05).

β (0.05 and 0.2) will result in fewer that 1 in 20 false positives only when the prior probability that there is a difference is in excess of approximately 50%. Hence, if the 'significant' result is a chance finding, there may be an unacceptably high probability of a false positive conclusion.

THE EFFECT OF UNDERPOWERED STUDIES

As previously noted, many published studies can be considered underpowered, and for these there is an increased chance that statistically significant results may, in fact, be false positive findings. Furthermore, many studies, even those with predefined hypotheses of interest, may also test numerous secondary hypotheses. Hence, even when there is a low prior probability that the principal null hypothesis is true, the prior probability that the secondary null hypotheses are true may actually be considerable. In such circumstances, there may be greatly increased risk that significant results for the secondary hypotheses are actually false positives. A recent study [16] observed that in 62% of the trials investigated, major discrepancies existed between the primary outcomes in the published reports compared to the original study protocols. In many of these there was evidence of a preference for statistically significant results to be published. Similarly for observational studies there is evidence of publication bias [4]. Hence, there is a substantial risk that the effects described here may combine with publication and reporting bias such that the (false) positive results of underpowered studies are published, but the true results of larger studies (i.e. where there was no effect) may be less likely to be published.

CONCLUSIONS

Whilst ethical issues associated with the increased risk of type II error with underpowered studies have been well documented, here we illustrate that such studies are also at increased risk of type I error, should the null hypothesis be rejected. The logic behind this statement can be illustrated through comparison of statistical testing with diagnostic test evaluation. There are many circumstances where the prior probability of the null hypothesis being true is high, not least of which arises during the analysis of secondary outcomes. These results indicate that statistically significant findings from studies with small sample sizes should be treated with increased scepticism, particularly where there is a reasonable chance that the null hypothesis is true.

REFERENCES

[1] Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. JAMA 1994; 272: 122-4.

[2] Freiman JA, Chalmers TC, Smith HJ, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of randomized controlled trials: survey of 71 "negative" trials. New Engl J Med 1978; 299: 690-4.

[3] Chan A-W, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. Lancet 2005; 365: 1159-62.

[4] Pocock SJ, Collier TJ, Dandreo KJ, *et al.* Issues in the reporting of epidemiological studies: a survey of recent practice. Br Med J 2004; 329: 883-7.

[5] Maggard MA, O'Connell JB, Liu JH, Etzioni DA, Ko CY. Sample size calculations in surgery: are they done correctly? Surgery 2003; 134: 275-9.

[6] Chung KC, Kalliainen LK, Spilson SV, Walters MR, Kim HM. The prevalence of negative studies with inadequate statistical power: an analysis of the plastic surgery literature. Plast Reconstr Surg 2002; 109: 1-6.

[7] Lochner HV, Bhandari M, Tornetta P. Type-II error rates (beta errors) of randomized trials in orthopaedic trauma. J Bone Joint Surg Am 2001; 83A: 1650-5.

[8] Altman DG. Statistics and ethics in medical research III: how large a sample size? Br Med J 1980; 281: 1336-8.

[9] Halpern SD, Karlawish JHT, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA 2002; 288: 358-62.

[10] Emanuel EJ, Wendler D, Grady C. What makes clinical research ethical? JAMA 2000; 283: 2701-11.

[11] Royall RM. The effect of sample size on the meaning of significance tests. Am Stat 1986; 40(4): 313-5.

[12] Oakes M. Statistical inference: a commentary for the social and behavioural sciences. Chichester: John Wiley & Sons; 1986.

[13] Neyman J, Pearson E. On the problem of the most efficient tests of statistical hypotheses. Philos T R Soc A 1933; 231: 289-337.

[14] Dawson B, Trapp RG. Basic and clinical biostatistics. 3rd ed. New York: McGraw-Hill Co.; 2001.

[15] Sterne JAC, Davey SG. Sifting the evidence - what's wrong with significance tests? Brit Med J 2001; 322: 226-31.

[16] Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials - Comparison of Protocols to published articles. JAMA 2004; 291: 2457-65.

Received: August 18, 2009 Revised: November 3, 2009 Accepted: December 10, 2009