大韓獸醫學會誌(2008) 第48卷 第3號
Korean J Vet Res(2008) 48(3) : 357~361

Sample size and statistical power consideration for diagnostic test research

Eu Tteum Kim[1], Choi Kyu Park[2], Son Il Pak[1,*]

[1]*School of Veterinary Medicine and Institute of Veterinary Science, Kangwon National University, Chunchon 200-701, Korea*
[2]*National Veterinary Research and Quarantine Service, Anyang 430-824, Korea*
(Accepted: September 19, 2008)

Abstract : Although power analysis is of important tool of research, investigators in veterinary medicine are unaware of the concepts of the statistical power. Two types of error occur in classical hypothesis testing and, those errors should be avoided, if possible. Since power is highly dependent on the sample size, whenever declaring non-statistically significant result they should consider the potential for committing a Type II error in their studies, which refers to the probability of falsely stating that two treatments are equivalent despite true difference between them. Also, sample size determination is one of the most important tasks facing the researcher when planning a diagnostic study, and provides valuable information on the characteristics of a test performance. This type of analysis forms the basis for proper interpretation of test results. The aim of this article was to re-evaluate some selected studies on diagnostic test reported in the domestic veterinary publications to determine the power and necessary sample size for inequality testing to ensure the desired power. Power calculations were illustrated using real-life examples of comparison of a new test and a reference test for detecting antibodies of various animal diseases. Factors affecting to the power were also discussed.

Keywords : diagnostic test, power, sample size, sensitivity, specificity

Introduction

Clinicians frequently perform a variety of tests as a diagnostic workup and then a decision is made to either rule-in or rule-out the presumptive diagnosis (hypothesis) based on test results. Whenever clinical decisions are made on the basis of the test results, two types of error can occur: a type I (α error, significance level) occurs if a researcher rejects a true null hypothesis; a type II (β error) occurs if a researcher fails to reject a false null hypothesis. That is, the former is the error of falsely stating that two tests are significantly different when there actually is none, whereas the latter is the error of falsely stating that two tests are equivalent when there is a difference. The power is the probability of rejecting correctly the hypothesis when the null hypothesis is false, and is referred to as $1-\beta$. In other words, it is the probability of saying there is a difference when a difference actually exists [1, 2, 5]. Because the relationship between alpha and beta error is inversely related each other, we could gain power by increasing the significance level, but we also increase the probability of making a false positive result. It is, therefore, important to maintain a balance between á error and power, depending on the specific research topics. Power analysis has been increasingly used in veterinary sciences [9, 14, 18] as well as in a variety of medical research fields [1, 10, 12, 16].

Evaluation of power of a study or a test must be a consideration during the initial planning of any research to determine the minimum number of animals that would be required not only to achieve specific objectives of interest but to satisfy sufficient power [9]. For ethical and economic reasons, the optimal sample size should not be so few as to miss biologically or clinically important effects or require unnecessary repetition of studies. This study was performed primarily to determine sample size required for evaluating diagnostic test characteristics of a new test in two situations: one is for comparing a new test with a

*Corresponding author: Son-Il Pak
School of Veterinary Medicine and Institute of Veterinary Science, Kangwon National University, Chunchon 200-701, Korea
[Tel: +82-33-250-8672, Fax: +82-33-244-2367, E-mail: paksi@kangwon.ac.kr]

reference test, the other for determining sensitivity and specificity of a new test when using reference samples. It was also aimed to assess statistical power of sample sizes which was reported in published articles.

Materials and Methods

Data and statistical analysis

For illustrative purposes, 8 articles were arbitrarily selected from the domestic veterinary publications [3, 4, 8, 13, 15, 17, 19, 20]. Sample size for paired-sample study was estimated using the following formula, given a desired power for detecting a one-sided meaningful difference (δ) with a one-sided test of size α [6, 7]: $n = [Z_\alpha \psi^{1/2} + Z_\beta(\psi - \delta^2)^{1/2}]^2 / \delta^2$, where $\delta = p1 - p2$; p1 and p2 for the sensitivity of a reference and a new test, respectively; ψ = probability of maximum (max) or minimum (min) disagreement between the two tests; $max = p1(1 - p2) + (1 - p1)p2$; $min = p1 - p2$; Z_α and Z_β for the value from the standard normal distribution corresponding to α and β, respectively. We calculated sample size for inequality testing using 80% of power and 5% significance level. Power for paired-sample design was initially calculated using the Microsoft Excel (Microsoft, USA) and then the value was validated with a freeware program, PS (The Netherlands). Results obtained from the software are presented.

Results

The number of sample size required for comparing sensitivity of a new test with a reference test, assuming 80% of power and one-sided 5% significance level is shown in Table 1, together with power of each study. For sample size calculation, the sensitivity of each test was cited either from the original paper or assumed values by the authors. The power of these studies when using maximum probability of disagreement ranged from a low of < 50% to a high of > 99%. For paired-sample design, the relationship between sample size and statistical power for 3 levels of significance level is shown in Fig. 1. For a given power, as the significance level increases, the sample size increases. Sample size estimates at an 80% of power combining significance levels and two types of hypothesis are shown in Fig. 2. Sample size of two-tailed test is increased compared to one-tailed test. For validating assay performance characteristics, the number of infected (or non-infected) samples required to achieve an expected sensitivity (or specificity) with a 2% allowable error of the estimates is shown in Fig. 3.

Discussion

The power of a test depends to a large extent on effect size, significance level, and sample size [5]. Others include type of hypothesis, variability in the population, and design effect [2, 9]. Effect size is a crucial parameter in power analysis, and is defined in several different ways. For practical reason, it is often defined as the minimum clinically important difference or change between groups. The larger the effect size,

Table 1. One-sided sample size estimates required for comparing sensitivity of various test kits with a reference test, assuming 80% of power and 5% significance level

Test or kit	No. samples tested	Reported or assumed sensitivity of a kit (%)	Sample size required*		Estimated Power (%)*	Reference
			Min	Max		
Neospra caninum	50	90.0	272	958	66.8	3
Coxiella burnetti	162	88.5	71	116	94.5	4
Toxoplasma	310	76.5	28	35	> 99.9	20
Bovine rhinotracheitis	96	64.5	17	20	88.2	13
Brucella	43	88.9	75	125	57.6	15
PRRS†	264	98.9	327	705	< 50.0	17
Avian influenza	242	95.5	398	1,881	> 99.9	8
Bluetongue	178	100.0	204	204	> 99.9	19

*For comparison purposes, sensitivity of reference test in each study was assumed to be 97%. All figures were not adjusted for prevalence rates of each disease. Sample size was estimated separately using minimum and maximum probability of disagreement. Power was calculated at one-sided 5% significance level using the reported or assumed values of sensitivity.
†PRRS, porcine reproductive and respiratory syndrome.

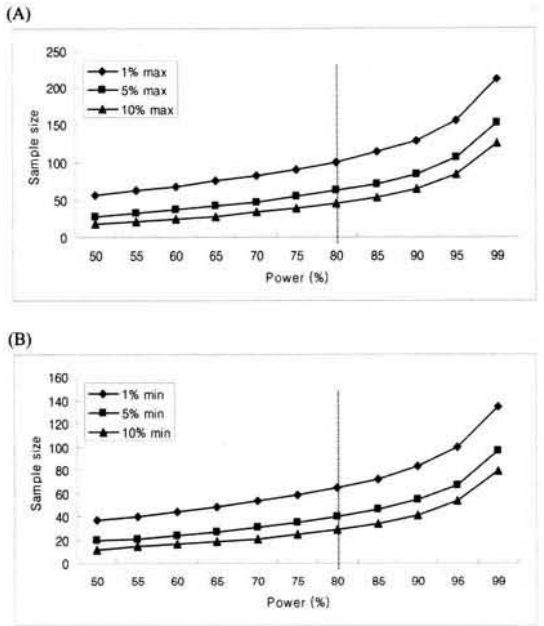

Fig. 1. Sample size required for a paired-sample design as a function of power using maximum (A) and minimum (B) probability of disagreement. This plot is for three significance levels of alpha = 1%, 5%, 10% (1-sided), and power ranging from 50 to 99% (1-sided). The dashed line represents 80% of power which is commonly used in biomedical research.

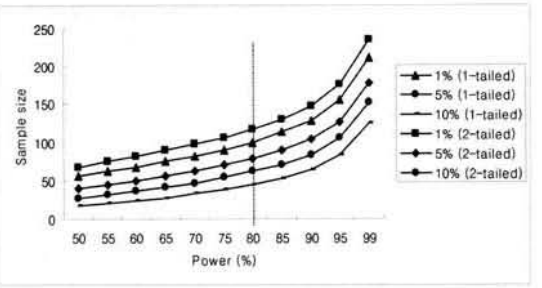

Fig. 2. Sample size estimates (maximum probability of disagreement) for a paired-sample design combining significance levels and two types of hypothesis. The dashed line represents 80% of power (1-sided) which is commonly used in biomedical research.

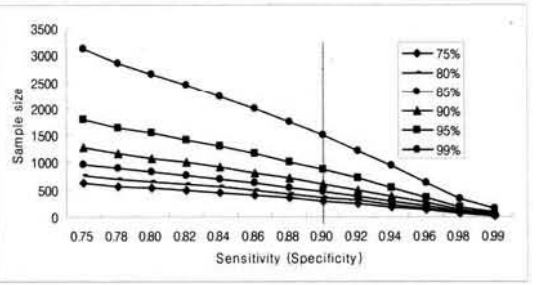

Fig. 3. Number of samples of known infection status required for validating a sensitivity or specificity, assuming allowable error of 2%. The dashed line is used in the example in the text.

the greater the power of the test. This means that for fixed sample sizes, as the effect size increases the p-value decreases. Similarly, the larger the effect size, the smaller the sample size necessary to detect it [5]. In this respect, studies with small effect size were found to have a relatively low power, as seen in Table 1.

The sample size is directly proportional to the power of a study: the larger the sample size, the study will have greater power to detect significant difference between the groups or an association between two variables. An underpowered study may not have a sufficiently large sample size to demonstrate the research question of interest, while an overpowered study may have too large sample size and wastes resources. In clinical trial and biomedical research, a minimum power of 80% is commonly used as acceptable level [5, 10]. With 80% of power, the type II error becomes 20%, indicating that the investigator has an 80% probability of achieving statistically significant differences or of rejecting a false null hypothesis. As noted previously, a type II error occurs when no statistically significant difference is detected between study groups. It does not necessarily mean that the effects of two groups are equal. The negative findings may be a true reflection of the lack of any difference between groups, or a result of insufficient sample size. With regard to the latter case, the results may be considered false negative result due to the probability of type II error being committed by the data as seen in every classical hypothesis testing [2, 9]. This type of error depends largely on sample size enrolled in the study. Therefore, the conclusion that there is no statistically significant difference between groups cannot be made unless a study has been proven to have sufficient power to detect difference when it exists.

An alpha error is inversely related to beta error (conventionally, beta error is set to 4 times alpha error) [5]. In comparison with two-tailed test, one-tailed test leads to a smaller sample size. Also, with all other things being equal including sample size, one-tailed tests usually are more powerful than two-tailed tests, indicating that the former may yield statistically

significant result more often that the latter might not [10, 14, 18]. For practical application of Fig. 1, suppose that a reference test has 80% sensitivity and that a researcher would like to detect an improvement of a new test with at least 90% of sensitivity at 80% of power and 5% of significance level with one-tailed test. In this case, the probability of disagreement ranges from a low of 10% to a high of 26%. The resultant number of samples at maximum and minimum probability of disagreement is 60 and 159, respectively.

Once the sample size required to achieve predefined acceptable power is obtained, then samples (animals) should be selected carefully to include as many as possible of those factors that may have an impact on outcomes. These may include breed, age, sex, stage of infection and others [12]. However, in many cases it is not feasible to fully represent all those variables in a finite population. In particular, assumption of normality which is basis for sample size calculation may not be satisfied when sample size is small. In this regard, it has been suggested that diagnostic study should include a minimum of 300 samples to provide confidence in the estimates of Se and Sp [12].

Estimates of sensitivity and specificity are among the important parameters to be obtained during validation of an assay because these values contribute to the basis for further calculation of other parameters. Considering gold test is rare for many animal diseases, and imperfect tests inherently yield false positive and false negative results, it is essential to acquire appropriate number and source of random samples to derive sensitivity and specificity. Because the sample size is inversely related to estimates of Se and Sp high estimates of those parameters will lead to insufficient sample sizes. For easy calculation of appropriate sample size for sensitivity (or specificity) researchers can use a nomogram by converting the values in Fig. 3 to tabulated one. The numbers calculated in Fig. 3 is based on a 2% allowable error of the sensitivity (or specificity) estimates. If researchers would modify the error, say 0.01, 0.04, 0.05, and 0.1, they only need to multiply the number of samples in the figure by a factor of 4, 0.25, 0.16, and 0.04, respectively. For instance, at a sensitivity of 90% with a 95% confidence level that the estimate is correct, the number of samples required is 865 at 2% allowable error. If we assume that 5% error is acceptable then the number of samples required is 139 (865×0.16). The sample size is dramatically reduced to 35 at 10% of error.

In conclusion, when designing any research project, sample size calculation should be made before initiation of a study to maximize statistical power, thus giving increased validity to the study. Further, we believe that editorial decision about publication should reflect the power of studies submitted, and for the underpowered studies with less than 50% the publication need to make that clear, regardless of its significance.

Acknowledgments

This study was supported by a grant (Project Code No., Z-AD21-2008-08-01) from National Veterinary Research & Quarantine Service, Ministry of Food, Agriculture, Forestry and Fisheries in 2008.

References

1. **Borm GF, Houben RM, Welsing PM, Zielhuis GA.** An investigation of clinical studies suggests those with multiple objectives should have at least 90% power for each endpoint. J Clin Epidemiol 2006, **59**, 1-6.
2. **Breau RH, Carnat TA, Gaboury I.** Inadequate statistical power of negative clinical trials in urological literature. J Urol 2006, **176**, 263-266.
3. **Cho Y, Kang S, Choi E, Jeong W, Yoon Y, Hwang W.** Development of indirect fluorescent antibody test and the prevalence of the antibody titer for *Neospora caninum* of domestic animal in Korea. Korean J Vet Res 1998, **38**, 595-599.
4. **Cho D, Kim Y, Wee S, Cho M, Kweon C, Kang Y, Park Y, Cho S.** Development of competitive enzyme linked immunosorbent assay for detection of *Coxiella burnetti* antibody in animal. Korean J Vet Res 2000, **40**, 81-85.
5. **Cohen J.** Statistical power analysis for the behavioral sciences. 2nd ed. pp. 52-56. Lawrence Erlbaum Associates Pub, Hillsdale, NJ, 1988.
6. **Connor RJ.** Sample size for testing differences in proportions for the paired-sample design. Biometrics 1987, **43**, 207-211.
7. **Dwyer AJ.** Matchmaking and McNemar in the comparison of diagnostic modalities. Radiology 1991, **178**, 328-330.
8. **Han M, Park K, Kwon Y, Kim J.** Comparison of

serological methods for detection of avian influenza virus antibodies. Korean J Vet Res 2002, **42**, 73-80 (in Korean).

9. **Hofmeister EH, King J, Read MR, Budsberg SC.** Sample size and statistical power in the small-animal analgesia literature. J Small Anim Pract 2007, **48**, 76-79.
10. **Houle TT, Penzien DB, Houle CK.** Statistical power and sample size estimation for headache research: an overview and power calculation tools. Headache 2005, **45**, 414-418.
11. **Huang W, LaBerge JM, Lu Y, Glidden DV.** Research publications in vascular and interventional radiology: research topics, study designs, and statistical methods. J Vasc Interv Radiol 2002, **13**, 247-255.
12. **Jacobson RH.** Validation of serological assays for diagnosis of infectious diseases. Rev Sci Tech 1998, **17**, 469-526.
13. **Jun M, Kim D, An S, Lee J, Min W.** Application of monoclonal antibody to develop diagnostic techniques for infectious bovine rhinotracheitis virus. II. Diagnosis of infectious bovine rhinotracheitis by using monoclonal antibody. Korean J Vet Res 1989, **29**, 27-35.
14. **Lenth RV.** Statistical power calculations. J Anim Sci 2007, **85**, E24-29.
15. **Lim Y, Lee D, Park J, Yang K, Kim S, Kim K, Hyun K, Kim W, Lee Y.** Enzyme-linked immunosorbet assay for detection of bovine antibody to *Brucella abortus*. Korean J Vet Res 1993, **33**, 131-135.
16. **Malik M, Hnatkova K, Batchvarov V, Gang Y, Smetana P, Camm AJ.** Sample size, power calculations, and their implications for the cost of thorough studies of drug induced QT interval prolongation. Pacing Clin Electrophysiol 2004, **27**, 1659-1669.
17. **Park C, Lyoo Y, Lee C, Jung J.** Comparison between indirect immunofluorescent antibody (IFA) test and enzyme-linked immunosorbent assay (ELISA) for the detection of antibody to porcine reproductive and respiratory syndrome virus (PRRSV). Korean J Vet Res 1998, **38**, 314-318.
18. **Roush WB, Tozer PR.** The power of tests for bioequivalence in feed experiments with poultry. J Anim Sci 2004, **82** (Suppl), E110-118.
19. **Shringi S, Shringi BN.** Comparative efficacy of standard AGID, CCIE and competitive ELISA for detecting bluetongue virus antibodies in indigenous breeds of sheep and goats in Rajasthan, India. J Vet Sci 2005, **6**, 77-79.
20. **Suh M, Joo H, Maass D.** Development of diagnostic kit (Test-MT) for the microplate latex agglutination test of toxoplasmosis in animal. Korean J Vet Res 1995, **35**, 583-593.