Family Practice
© Oxford University Press 1997

Vol. 14, No. 4
Printed in Great Britain

Empowering research: statistical power in general practice research

Nick Fox and Nigel Mathers

Fox N and Mathers N. Empowering research: statistical power in general practice research. *Family Practice* 1997; **14**: 324–329.

Background. Statistical power is a measure of the extent to which a study is capable of discerning differences or associations which exist within the population under investigation, and is of critical importance whenever a hypothesis is tested by statistics. Conventionally, studies should reach a power level of 0.8, such that four times out of five a false null hypothesis will be rejected by a study. Statistical power may most easily be increased by increasing sample size.

Objective. We aimed to assess the level of statistical power of general practice research.

Methods. A total of 1422 statistical tests in 85 quantitative original papers in the *British Journal of General Practice* were analysed for statistical power.

Results. The median power of tests analysed was 0.71, representing a slightly greater than two-thirds likelihood of rejecting false null hypotheses. Of 85 studies, 37 (44%) attained power of 0.8 or more. Ten studies had power of more than 0.99 suggesting 'over-powering'. Twenty-one of the papers surveyed (25%) had a likelihood of gaining significant results poorer than that obtained by tossing a coin when a null hypothesis is false.

Conclusion. While achieving higher power than studies in similar surveys of other disciplines, the power of general practice research falls short of the 0.8 convention. Adequate power is essential so that effects which exist are not missed. Recommendations are made concerning power calculations prior to the start of research and reporting of results in journal articles.

Keywords. General practice research, research publication, sampling, statistical power.

Introduction

In quantitative research, decisions about sample size are crucial. Purely at the level of resource costs, the size of sample has an impact: achieving larger sample sizes generally entails increased costs in terms of staffing for data collection and analysis, and other research expenses. In addition, certain research topics will impose their own constraints about the size of sample which can realistically be achieved. For example, a study based upon a single GP patient population may face diminishing numbers of respondents who may be recruited when a condition is relatively uncommon or a particular treatment regimen is undertaken. Even where an initial sample size is quite large, subsequent stratifications by factors such as gender and age may soon reduce the sample to small numbers.

The objective of tests of statistical inference is to generalize study findings to the population under investigation. Statistical tests of differences between groups or associations between variables seek to disprove a null hypothesis (H_0) which states that there is no difference or no association within the population. If a null hypothesis can be rejected, it means that some difference or some association may be inferred from the study to the wider population. However, the potential for a statistical test to disprove a null hypothesis depends upon the power of the study, and as such, statistical power is crucial to quantitative research. The need for a study to possess adequate power can be illustrated by considering the four situations which can arise when the data from a study are analysed using

Received 16 December 1996; Accepted 3 April 1997.
Institute of General Practice and Primary Care, School of Health and Related Research, University of Sheffield. Correspondence to N Fox at Community Sciences Centre, Northern General Hospital, Herries Road, Sheffield S5 7AU, UK.

		POPULATION	
	Null Hypothesis is:	False	True
STUDY	False	Correct Result	Type I Error (alpha)
	True	Type II Error (beta)	Correct Result

FIGURE 1 *The null hypothesis (H_0) statistical significance and statistical power*

statistical tests of inference. These are illustrated in Figure 1.

Each cell in the figure represents a possible relationship between the findings of the study and the 'real-life' situation in the population under investigation (but which of course cannot be known other than through statistical inference). Cells 1 and 4 represent desirable outcomes, while cells 2 and 3 represent potential outcomes of a study which are undesirable and need to be minimized.

Cell 1

The null hypothesis has been disproved by the results of the study (that is, there is support for a hypothesis which suggests some differences between groups or association between variables). This is also the situation in the population. Thus, we can be satisfied that the study is reflecting the world outside the limits of the study and it is to be accepted as a 'correct' result.

Cell 4

The results from the study support the null hypothesis. This is the situation which pertains in the population, so we can be satisfied that our study reflects the circumstances in the population. Once again, this is a 'correct' result.

Cell 2

In this cell, as in cell 1, the study results falsify the null hypothesis, indicating some kind of difference or association between variables. However, in the world beyond the study, the null hypothesis is actually true and there is no effect. This is known as a Type I error: the error of wrongly rejecting a true null hypothesis. The likelihood of committing a Type I error is known as the alpha (α) value or the statistical significance of the test. Many readers will be more familiar with alpha as the quoted *P* level of significance of a test. The *P* value marks the probability of committing a Type I error; thus a *P* value of 0.05 indicates a 5% (or 1 in 20) chance of committing a Type I error. Cell 2 thus reflects an incorrect finding from a study, and the alpha value represents the likelihood of this occurring.

Cell 3

This cell similarly reflects an undesirable outcome of a study. Here, as in Cell 4, a study supports the null hypothesis, implying that there is no difference or association in the population under investigation. But in reality, the null hypothesis is false and there is some kind of difference or association which the study is missing. This mistake is known as a Type II error and is the error of wrongly accepting a false null hypothesis. The likelihood of committing a Type II error is the beta value of a statistical test, the complement ($1-\beta$) is the statistical power of the test. Thus, the statistical power of a test is the likelihood of avoiding a Type II error when the null hypothesis is false. Conventionally, a value of 0.80 or 80% is the target value for statistical power, representing a likelihood that four times out of five a false null hypothesis will be correctly rejected. Outcomes of studies which fall into cell 3 are incorrect; β, or its complement power are the measures of the likelihood of such an outcome of a study.

All research should seek to avoid both Type I and Type II errors, which lead to incorrect inferences about the world beyond the study. In practice, there is a trade-off. Reducing the likelihood of committing a Type I error by increasing the level of significance at which one is willing to accept a positive finding reduces the statistical power of the test, increasing the possibility of a Type II error, and vice versa. However, both statistical significance and statistical power are affected by sample size. While researchers are usually aware that the chances of gaining a significant result will depend on the size of a study's sample, explicit power calculations are often not undertaken prior to the start of a study, and the evidence from various fields of study is that many studies do not meet the 0.8 conventional target for power.[1-4] Such under-powered studies have much reduced likelihoods of being able to discern the effects which they set out to seek: a study with a power of 0.66 will only detect an effect two times out of three, while studies with power of 0.5 or less will detect effects at levels less frequent than those achieved by tossing a coin. A non-significant finding of a study may thus simply reflect the inadequate power of the study to detect differences or associations at levels which are conventionally accepted as statistically significant.

Statistical power calculations can be undertaken after a study has been completed, and this paper reports such analyses on general practice research. More importantly, such calculations need to be undertaken prior to a study to avoid both the wasteful consequences of under-powering, and of overpowering in which sample sizes are excessively large, leading to very high power at the expense of higher than necessary study costs.

Power is a function of three variables: the level of significance (α), the effect size (the measure of 'how wrong' a null hypothesis is),[4] and the sample size. While calculation of power entails recourse to tables

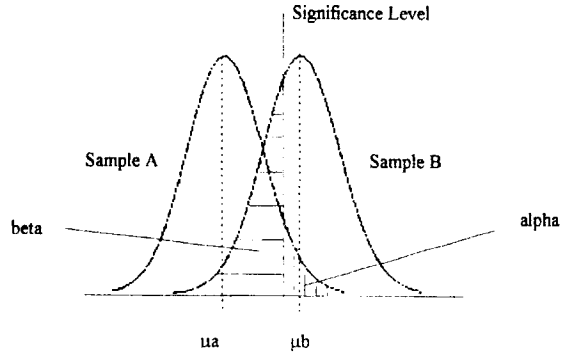

FIGURE 2 *The risk of Type I and Type II errors for* $\alpha = 0.05$

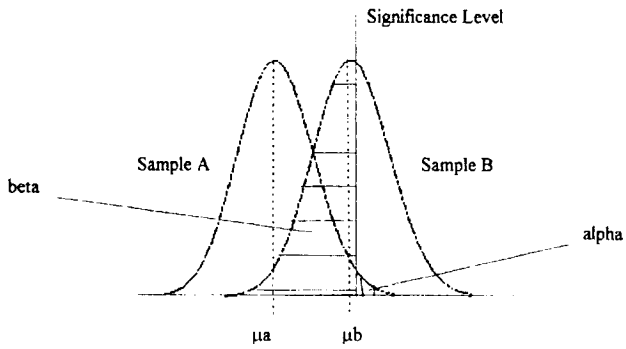

FIGURE 3 *The risk of Type I and Type II errors for* $\alpha = 0.01$

of values for these variables, the calculation is relatively straightforward in most cases.

Firstly, power depends on the level of statistical significance chosen by the researchers. The more stringent the level of significance chosen, the lower the power. The trade-off between significance and power is illustrated by Figures 2 and 3 which represent the distributions of two differing samples A and B, for instance the heights of female and male adults. If the mean of population A is μa and that of population B is μb, then the null hypothesis that there is no difference between population means, H_0: $\mu a = \mu b$, while the hypothesis that the mean of B is greater than that of A, H_1: $\mu b > \mu a$.

In Figure 2, the one-tailed α value of 0.05 has been selected and is represented by the vertical cut-off line. By drawing this line we are saying that all scores to the right of this line come from sample B. In this example, we are likely to be right much of the time. However, the scores to the right of this cut-off line include the shaded portion containing 5% of sample A scores. At this level of significance, 5% of sample A is likely to be mistakenly allocated to sample B. This is the chance of a Type I error, of concluding that the sample comes from B when it does not. In addition, the shaded area to the left of the cut-off represents scores from sample B which do not achieve the desired significance level, and thus are mistakenly allocated to

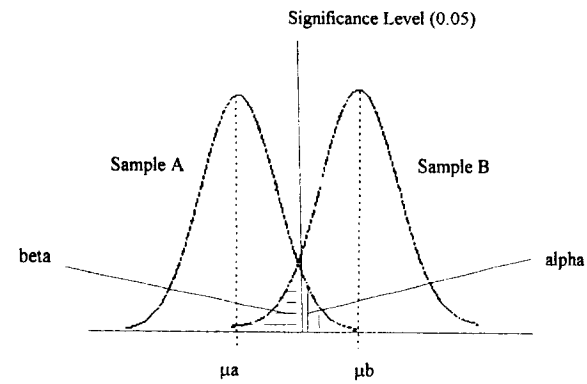

FIGURE 4 *Impact on power of increasing effect size*

sample A. This portion represents the β value, that is, the likelihood of missing a real difference between the two distributions, of committing a Type II error.

In Figure 3, the α value has been set at a value of 0.01, and as can be seen, the shaded area to the right of the line is smaller, while that to the left is substantially larger. Thus as α is reduced β increases, and the concomitant power $(1-\beta)$ is lower, Researchers need to decide the relative consequences of committing a Type I error (for example, missing a real improvement in management of a condition supplied by a new but costly treatment) and committing a Type II error (for instance, missing subtle side effects of a proposed new drug), and set the α value accordingly. If the α value is more stringent sample size will need to be increased to compensate. If the directionality of a hypothesis can be stated then it is legitimate to use one-tailed as opposed to two-tailed tests (where possible) and this will increase power.

The second factor is the effect size (ES) which is under investigation in the study. Except in research which is effectively repeating earlier studies, the size of an effect will not be known, as it reflects the very population variation which a piece of research sets out to study. However it is essential that estimates of ES are made, because unless an effect size is large then many studies with small sample sizes are likely to be under-powered. Once again this can be illustrated graphically. Figure 2 demonstrates a relatively small effect between two distributions for which a *t*-test might be used to explore the extent of difference. In Figure 4, the α value remains at 0.05, but now the ES (reflected in the difference between the means of the two distributions) is larger. As can be seen, the shaded area representing β is much reduced, and the power increased.

Accurate estimation of effect sizes is essential for calculating power before a study begins, particularly where multiple hypotheses are being tested. It is sometimes possible to increase effect size, but usually this is the intractable element in the equation. So researchers need to consider how large sample sizes

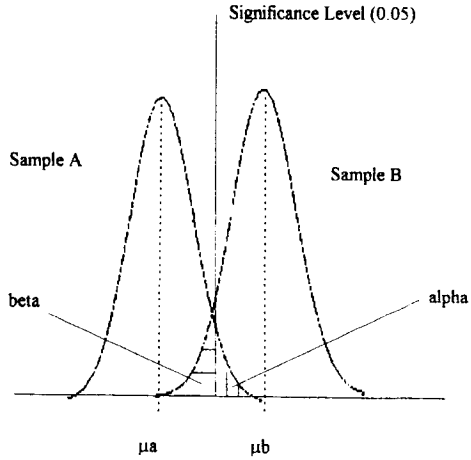

FIGURE 5 *Impact on power of increasing sample size*

will need to be to adequately test all the hypotheses in a study, and this assessment will affect considerations of research design. The effect of increasing the size of a sample is illustrated in Figure 5. While the effect size ($\mu a - \mu b$) remains the same as in Figure 2, the variance (represented by the area under the curves) is reduced and, for an α of 0.05, β is reduced.

An ES can be estimated in four ways:[4] first, from a review of the literature or meta-analysis, which can suggest the size of ES which may be expected; second, by means of a pilot study which can gather data from which the size of effect may be estimated; third, one can make a decision about the smallest size of effect which it is worth identifying—for example, if we wish to assess the relative effectiveness two rival drugs, if we are willing to accept the two drugs as equivalent if there is no more than a 10% difference in their efficacy, then this effect size may be set, acknowledging that smaller effects will not be discernible. Finally, as a last resort, one can use a 'guesstimate' as to whether an ES is 'small', 'medium', or 'large'. Definitions and values for 'small', 'medium' and 'large' effects for a range of statistical tests, based on figures in Ref. 5, are set out in Table 1, along with formulae for calculating these effects for each test. A 'medium' effect is defined as one which is 'visible to the naked eye'—in other words, one which could be discerned from everyday experience without recourse to formal measurement. For example, the difference between male and female adult heights in the UK would be counted as a medium ES. Using the formula in Table 1 for calculating an ES for a *t*-test, if mean heights for women and men are 160 and 175 cm respectively, and the pooled standard deviation is 7.5 cm, the ES = 175 - 160/7.5 = 0.5.

Most effects encountered in biomedical and social research should be assumed to be small, unless there is a good reason to claim a medium effect, while a 'large' effect size would probably need to be defined as one which is so large that it hardly seems necessary to undertake research into something so well established. Cohen offers the example of the difference between the heights of 13- and 18-year-old girls as a 'large' effect.[5]

Power calculations may also be used as part of the critical appraisal of research papers. Unfortunately it is rare to see β values quoted for tests in research reports; indeed, the results reported are often inadequate to calculate effect sizes. In the last part of this paper, we offer recommendations to journal editors and contributors on how results should be presented to enable readers to assess the power of a study. Formulae for calculating effect sizes vary from test to test (see Table 1), but most are relatively easy to calculate, and more information can be obtained from various texts on the subject.[5,6] Appraisal of study power can be particularly difficult where negative results are simply reported as 'not significant' without supplying any details of test statistics, raw results or degrees of freedom. Statistically non-significant results can be highly significant clinically. However, if a study fails to reject a null hypothesis, it is important to know whether this is because the null hypothesis is correct or whether the test is simply under-powered and thus unlikely to produce statistically significant results.

TABLE 1 *Effect size definitions[a] and formulae*

Test	'Small'	'Medium'	'Large'	Formula
t-test	0.2	0.5	0.8	$d = \mu a - \mu b/\sigma$
F test	0.1	0.25	0.4	σ of means/pooled σ
Correlation (Pearson)	0.1	0.3	0.5	r
χ^2	0.1	0.3	0.5	$(\chi^2/N)^{-1}$
Multiple regression	0.02	0.15	0.35	$(R^2/1-R^2)^{-1}$

[a] Effect sizes for non-parametric tests may be assumed to be as for their parametric equivalents.
μ_a = mean of sample a; σ = standard deviation; r = correlation coefficient; χ^2 = chi square test statistic; R^2 = multiple R squared.

TABLE 2 *Types of research paper published in period January 1994–June 1995*

Type of paper	N
Qualitative	7
Meta-analysis or discursive	4
Quantitative without hypothesis testing	35
Quantitative: non-standard tests	6
Quantitative	85
Total	137

TABLE 3 *Power of studies* (n = *85)*

Power band	N	%
<0.25	2	2
0.26–0.49	19	22
0.50–0.79	27	32
0.80–0.96	21	25
≥0.97	16	19

Survey of statistical power in general practice research

To explore the power of general practice research, we analysed all the statistical tests reported in the *British Journal of General Practice* (BJGP) over a period of 18 months. Power was calculated for each test based on the reported sample size. This enabled calculation of the power of each quantitative study published during this period, to assess the adequacy of sample sizes to supply sufficient power.

Methods

All original research papers published in the BJGP during the period January 1994 to June 1995 inclusive were analysed in terms of the power of statistical tests reported. Table 2 indicates the breakdown of papers by type. Qualitative papers were excluded, as were meta-analyses and articles which, although reporting quantitative data, did not report any formal statistical analysis even though in some instances such tests could have been undertaken. A further six papers were excluded because they did not use standard statistical tests for which power tables were available. This left 85 papers, involving 1422 tests for which power could be calculated using power tables.[5,6] Power was calculated for each test following conventions of similar research into statistical power.[1–4] Where adequate data were available (for example details of group means and standard deviations, or results of χ^2 tests) precise effect sizes could be calculated. Where this was not possible (in particular for results simply reported as 'non-significant') the following assumptions were made, all of which considerably over-estimate the power of the test. First, for significant results, the effect size was assumed to be 'medium', which as noted earlier means an effect 'visible to the naked eye'. Non-significant results were assumed to have a 'small' ES. Second, α values were set at the lowest possible conventional level of 0.05, and where a directional test was used, a one-tailed α was used (equivalent to two-tailed α of 0.1).

From the calculations of power for individual tests, a mean power for each paper was derived. This strategy has been adopted in other studies into statistical power:[1–4] what is reported is study power, rather than test-by-test power, providing an estimate of the quality of studies in terms of overall adequacy of their statistical power.

Results

Eighty-five papers comprising 1422 tests were analysed. The median number of tests per paper was 12, with a minimum of one test and a maximum of 90. The median power of the 85 studies was 0.71, representing a slightly greater than two-thirds probability of rejecting null hypotheses. The proportions of tests in different power bands is summarized in Table 3. Of the 85 studies, 37 (44%) had power of ≥0.8, while 48 (56%) fell below this conventional target. The lowest power rating was 0.24, while 10 studies (12%) reached power values of 0.99 or more. Unlike some earlier studies of statistical power, no attempt has been made to subdivide studies into those with large, medium and small effect sizes, partly because in a number of cases, the effect sizes were probably over-estimated as 'medium', and because within a single paper, many different variables with differing effect sizes might be under investigation.

Discussion

The results of this survey of general practice research published in the BJGP indicates somewhat higher power ratings than those reported for other disciplines, including nursing,[4] psychology,[2] education,[1] management[3] and in some medical journals.[7] However, over half of the studies fell below the conventional figure of 0.8, and 25% had a power of 0.5 or less, suggesting a chance of gaining significant results poorer than that obtained by tossing a coin.

Scrutiny of the distribution of powers indicated bimodality. Sixteen of the 37 papers meeting or exceeding the 0.8 target had a power of >97%. Such high powers were achieved by the use of very large samples. Given that it is necessary to double the sample size to increase power from 0.8 to 0.97, it is reasonable to argue that as such the studies were overpowered, and used sample sizes which were excessively expensive

in terms of researcher time for data collection and analysis. In some cases these studies used pre-existing data sets and so this criticism is less pertinent; elsewhere, researchers may have devoted far greater efforts in terms of time and obtaining goodwill from subjects than may strictly have been necessary to achieve adequate power. The importance of pre-study calculations of necessary sample size to achieve statistical power of 0.8 or thereabouts is relevant both for those studies demonstrated to be under-powered and those for whom power is excessive.

Conclusions and recommendations

More than half of the published quantitative papers in general practice research surveyed in this study possessed inadequate statistical power. This means that during the statistical analysis there was a substantial risk of missing significant results. Twenty-five per cent of papers surveyed had a chance of gaining significant results (when there was a false null hypothesis) poorer than that obtained from tossing a coin. With regard to the use of statistical power analysis in general practice research, we would recommend that all primary care and general practice researchers should undertake power calculations, referring to relevant texts or to a statistician, to decide on the necessary sample size before starting research. Furthermore, we would tactfully suggest to editors of general practice journals that they should request authors to report the value of each test statistic with the α and β values as well as sample size, to enable readers to assess the power of a study. All such values for non-significant results should also be reported.

References

[1] Brewer J. On the power of statistical tests in the American Educational Research Journal. *Am Educ Res J* 1972; **9:** 391–401.

[2] Chase J, Chase B. A statistical power analysis of applied psychological research. *J Appl Psychol* 1976; **61:** 234–237.

[3] Mazen AM, Graf LE, Kellog CE, Hemmasi M. Statistical power in contemporary management research. *Acad Man J* 1987; **30:** 369–380

[4] Polit DF, Sherman RE. Statistical power in nursing research. *Nurs Res* 1990; **39:** 365–369

[5] Cohen J. *Statistical Power Analysis for the Behavioural Sciences*. New York, NY: Academic Press, 1977.

[6] Machin D, Campbell MJ. *Statistical Tables for the Design of Clinical Trials*. Oxford: Blackwell Scientific, 1987.

[7] Reed JF, Slaichert W. Statistical proof in inconclusive 'negative' trials. *Arch Intern Med* 1981; **141:** 1307–1310.