Is the p value really so significant?*

Arianne P Verhagen[1], Raymond WJG Ostelo[2,3] and Arno Rademaker[4]

[1]Department of General Practice, Erasmus MC, University Medical Center Rotterdam [2]VU University Medical Centre, EMGO Institute

[3]Research Group of the Amsterdamse Hogeschool voor Paramedische Opleidingen [4]University of Professional Education Brabant, Hogeschool van Brabant

In the past decade attention has been paid to ‘evidence-based practice’ which seeks to demonstrate the effectiveness of treatment strategies with evidence from randomised studies. Statistical techniques are used to draw conclusions from the results of such research. A frequently applied statistical measure is the p value. This research note describes advantages and disadvantages of the *p* value in randomised studies.

The most appropriate design to study the effect of treatment is the randomised clinical trial. In the simplest example of a randomised trial, two treatments (e.g., massage and exercise) are compared in patients with a specific complaint. The crucial question the researchers wish to answer is whether one treatment is more effective than the other.

The *p* value is a statistical variable that, according to many people, provides an answer to the question about whether the difference in effect between two treatments depends on chance. This procedure is called testing for statistical significance. Within medical science there is great belief in significance testing as a method of analysis, but (over) appreciation of the *p* value has been the subject of criticism for several years (Nurminen 1997).

What is the p value?

In statistics the term ‘significance’ is used to mean that a certain study result is not a chance finding, and that there really is something going on. Statisticians always start with the assumption that a study result might depend on chance. Therefore the basic proposition in every statistical test is the null hypothesis, which represents the theory that both treatments (massage and exercise) are equally effective. When the null hypothesis is true, outcomes differ only by chance.

Assume that a randomised trial has been set up to study the effect of exercise compared to massage. The primary outcome measure is ‘perceived recovery’ (on a 7-point Likert-scale patients indicate the extent to which they have recovered, ranging from ‘total recovery’ to ‘worse than ever’). We could consider every patient who indicates slight to total recovery to be recovered, whereas all others have not recovered. With these outcomes it is possible to calculate whether more patients have recovered with exercises than with massage.

It is usual to test the outcome statistically. As a result of the natural course of the condition most patients will recover to a certain extent anyway, and there will also always be some difference in recovery between the two groups by chance. What is tested is whether the difference between the two groups is greater than can be expected based on chance.

The null hypothesis in this specific case is: ‘There is no difference between the effect of exercise and massage.’ The alternative hypothesis is therefore: ‘The difference in recovery between the two groups is greater than that which could be expected on the basis of chance.’ Thus, the effect of exercises could be *greater or smaller* that the effect of massage. This is referred to as the *two-tailed* testing of the null hypothesis.

The problem with statistical tests is that there is always a risk, even if the null hypothesis is correct, that chance (incorrectly) suggests that the alternative hypothesis is true. To return to the example: assume that there really is no difference in effect between exercises and massage on the recovery of patients, and yet it is found that more patients who received massage recover than the patients in the exercise group. In that case the null hypothesis is incorrectly rejected (a ‘Type I’ error). This risk is indicated with the *p* value or α. Whenever you see ‘$p < 0.05$’, this means that when the null hypothesis is rejected and the alternative hypothesis is accepted, there is a risk of less than 5% that this decision is incorrect.

In scientific research it is considered (relatively arbitrarily) that when the *p* value that is found is less than 5% it is acceptable to reject the null hypothesis. (Another way of saying this is that α, the critical *p* value, is set at 0.05.) In that case, when $p < 0.05$ the finding is considered statistically significant. The fact that α is almost always set at 5% is one of the points of criticism. Situations can occur in which a 10% risk of an incorrect decision (null hypothesis incorrectly rejected) is also appropriate, or when this risk must be kept as low as possible and should be set at 1%.

Significant versus non-significant

Testing for statistical significance appears to be an objective way to determine whether a null hypothesis should be rejected (Nurminen 1997). The *p* value is often used as a dichotomous measure of evidence: the *p* value is less than/greater than 0.05; so the finding is significant/non-significant, resulting in the conclusion that the treatment is/is not effective.

Assume that a value of $p = 0.049$ is found. The null hypothesis (in our example, the hypothesis that exercise is as effective as massage) is rejected. The alternative hypothesis is therefore accepted, which is often interpreted as indicating that the intervention (exercise in this example) is effective. If, instead, the value were $p = 0.055$, the null hypothesis would not be rejected because the difference in effect between exercise and massage is not statistically significant.

Technically it is incorrect to interpret a *p* value of 0.055 as evidence that the intervention is not effective, because ‘No evidence of effect is not evidence of no effect’ (Altman and Bland 1996). Apart from a considerable number of methodological problems that can be encountered in clinical trials, from a statistical point of view there is one important reason why it may not be possible to demonstrate the effect of a treatment, namely ‘lack of power’. The *p* value is not only based on the difference that is found between two treatments, but also depends on the number of patients in the two groups (Goodman 1999). A small difference in treatment effect between the intervention and control groups in a study with many patients (e.g. 10 000) can produce the same *p* value as a large difference in effect between the two groups in a study with only a few patients (e.g. 50). When there are too few patients in the treatment groups to be able to demonstrate a difference this is called a 'Type II' error.

Validity

The *p* value only says something about whether the null hypothesis should be rejected, but nothing about the validity of the null hypothesis (Slakter, Wu and Suzuki-Slakter 1991). Assume that the null hypothesis is actually correct. Then, if α is set at 0.05 a significant chance result will be found in less than 1 in 20 cases. If the same null hypothesis were tested more than 20 times (if the same clinical trial were carried out more than 20 times), a 'significant' result would probably be found at least once. This one time could, in fact, be in the trial that is being carried out at that moment.

The same 'game of chance' can also occur within a study, namely in a study in which more than 20 outcomes are measured. Based on chance, one of these outcome measures will probably be 'significant'. Adjustments must be made in the calculations for this 'game of chance.' However, such adjustments are often not made and researchers claim, for instance, that 'their' intervention is effective based on the effect measure that reaches statistical significance, whereas in fact it is only one of the 20 effect measures that they have included in their study.

Clinical relevance

A statistically significant result is not necessarily clinically relevant. Imagine that a randomised trial has been carried out in 4000 patients. The most important outcome is pain measured on a 100 mm visual analogue scale (VAS). After randomisation it appears that the two groups are comparable with regard to the average score for the level of pain: both the intervention group and the control group scored 75 mm on the VAS. At the end of the treatment the average level of pain in the control group has dropped to 45 mm and in the intervention group to 40 mm. To the great joy of the researcher, this difference is significant, and the conclusion is drawn that exercise is more effective than massage. But is the intervention really so effective? Is an average difference of 5 mm on the VAS between the two study groups really an indication of significantly less pain? In both groups there still is a considerable amount of pain.

In order to be able to say something about the clinical relevance of the effect that is found, it is better if the researcher indicates at the start of the study that a minimal difference of 10 or 20 millimetres on the VAS will be considered clinically relevant.

Alternatives to p values

If a study could be carried out hundreds of times, the same results would not be found each time. On average, a certain amount of difference can be found between the two treatment groups. In 95% of the studies that are carried out the average difference that was found would lie between certain values on either side of the estimated 'real' difference in treatment effect. These values are called the '95% confidence limits' and the region they enclose is called the '95% confidence interval'.

When a study is carried out only once, how do we know that the difference between the two treatment groups found is anywhere near the 'real' difference? The answer is that we don't know. If a 95% confidence interval is calculated for the result that is found, then this can be interpreted as follows. Suppose the trial was replicated a hundred times, the 95% confidence interval would be wide enough that 95% of the trials would yield an estimate within this confidence interval.

Assume that it is found that there is a difference of 5% between the intervention and control group, and the hypothetical 95% confidence interval surrounding this difference ranges from -2% to 12%. The conclusion might be that the 95% confidence interval also contains 0% difference, so the true effect could be no effect. This is equivalent to a non-significant result; the null hypothesis is not rejected. According to Guyatt et al (1995) it is also possible to say that there is probably a real difference in treatment effect that is nearer to 5% than to -2% or 12%. A reasonable conclusion of this clinical trial might be that there is a non-significant outcome and only a small treatment effect.

References

Altman DG and Bland JM (1996): Absence of evidence is not evidence of absence. *Australian Veterinary Journal* 74: 311.

Goodman SN (1999): Towards evidence based medical statistics 1: The p value fallacy. *Annals of Internal Medicine* 130: 995–1004.

Guyatt G, Jaenschke R, Heddle N, Cook D, Shannon H and Walter S (1995): Basic statistics for clinicians 2. Interpreting study results: Confidence intervals. *Canadian Medical Association Journal* 152: 169–173.

Nurminen M (1997): Statistical significance; A misconstrued notion in medical research. *Scandinavian Journal of Work Environment and Health*. 23: 232–235.

Slakter MJ, Wu YB and Suzuki-Slakter NS (1991): *, **, and ***; Statistical nonsense at the .00000 level. *Nursing Research* 40: 248–249.

Further reading

Craen AJM de, Vickers AJ, Tijssen JGP and Kleijnen J (1998): Number-needed-to-treat and placebo controlled trials. *Lancet* 351: 310.

Greenhalgh T (1997): Statistics for the non-statistician II: 'Significant' relations and their pitfalls. *BMJ* 315: 422–425.

This research note has previously been published as: Is de p-waarde wel significant (2000). Nederland Tijdschrift voor Fysiotherapie 110: 92–95; and as a chapter in Ostelo RWJG, Verhagen AP and de Vet HCW (Eds) (2002) *Education in Scientific Research* [Dutch]. Houten, Bohn Stafleu van Loghum.