Scandinavian Journal of Pain 4 (2013) 220–223

Contents lists available at ScienceDirect

Scandinavian Journal of Pain

journal homepage: www.ScandinavianJournalPain.com

Topical review

Significance tests in clinical research—Challenges and pitfalls

Eva Skovlund[a,b,*]

[a] School of Pharmacy, University of Oslo, Norway
[b] Norwegian Institute of Public Health, Oslo, Norway

HIGHLIGHTS

- Statistical analysis of data must be carefully planned before a clinical trial is initiated.
- Expected effect size and estimates of necessary sample size are important parts of trial-planning.
- *p*-values do not assess the size of an effect.
- *p*-values do not tell us whether a statistically significant result is of any clinical relevance.
- Effect estimates and corresponding 95% confidence intervals should always be reported.

ARTICLE INFO

Article history:
Received 19 July 2013
Accepted 19 July 2013

Keywords:
Statistical analyses
Clinical trials
p-values
Effect estimates
Confidence intervals

ABSTRACT

Background: Statistical analyses are used to help understand the practical significance of the findings in a clinical study. Many clinical researchers appear to have limited knowledge on how to perform appropriate statistical analysis as well as understanding what the results in fact mean.
Methods: This focal review is based on long experience in supervising clinicians on statistical analysis and advising editors of scientific journals on the quality of statistical analysis applied in scientific reports evaluated for publication.
Results: Basic facts on elementary statistical analyses are presented, and common misunderstandings are elucidated. Efficacy estimates, the effect of sample size, and confidence intervals for effect estimates are reviewed, and the difference between statistical significance and clinical relevance is highlighted. The weaknesses of *p*-values and misunderstandings in how to interpret them are illustrated with practical examples.
Conclusions and recommendations: Some very important questions need to be answered before initiating a clinical trial. What is the research question? To which patients should the result be generalised? Is the number of patients sufficient to draw a valid conclusion? When data are analysed the number of (preplanned) significance tests should be kept small and *post hoc* analyses should be avoided. It should also be remembered that the clinical relevance of a finding cannot be assessed by the *p*-value. Thus effect estimates and corresponding 95% confidence intervals should always be reported.

Contents

DOI of refers to article: http://dx.doi.org/10.1016/j.sjpain.2013.08.001.

[*] Corresponding author at: School of Pharmacy, University of Oslo, Norway. Tel.: +47 22 85 61 72; fax: +47 22 85 44 02.
E-mail address: eva.skovlund@farmasi.uio.no

1877-8860/$ – see front matter © 2013 Scandinavian Association for the Study of Pain. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sjpain.2013.07.023

1. Introduction

In the medical literature significance tests are extensively used as a method of comparing the effect of different treatments or groups based on patient characteristics. Traditionally a low *p*-value ($p < 0.05$) is interpreted as proof of an interesting effect, whereas a higher *p*-value is interpreted as no effect of the variable under study. That might be a valid conclusion, but it is not necessarily so, and one might question whether this is actually sound practice.

2. Methods

This is a mini-review of common, but often not well understood challenges clinicians meet when analyzing data and assessing both statistical significance and clinical relevance. The review is based on long experience in conducting clinical research and broad experience as a consultant in medical statistics for medical journals. Explanations and advice to novices as well as experienced authors of health science manuscripts will follow.

3. Significance tests, confidence intervals, effect size, sample size, and *p*-values

First of all, the p-value is the result of a significance test and answers a specific question: What is the probability of observing the actual estimated effect or an even larger one if the null hypothesis is true? Since the most common null hypothesis is that two groups or treatments are not different, the question could be simplified as follows: What is the probability that the observed (or a larger) effect is observed *just by chance*?

The general formula of a (parametric) test statistic *t* is

$$t = \frac{O - E}{S}$$

where O is the observed value of the outcome variable (for instance a proportion or mean), E the expected outcome under the null hypothesis, and S a measure of random variation (standard error). Putting it simple, if there is no effect of treatment, the difference $O - E$ and hence t, should be close to 0.

As can be seen, the *size of the effect* ($O - E$) is measured relative to random variation. The random variation will decrease with an increasing number of observations (n). Given that the effect is unchanged, increasing n would lead to a larger value of t. The test statistic t is converted to a *p*-value. Generally, the larger the absolute value of t, the lower the *p*-value and correspondingly the probability that an effect is due to chance. Thus, *a p-value does neither assess the size of an effect as such nor whether a statistically significant result is of any clinical relevance.* In order to obtain clinically relevant information, the size of the effect as well as a *95% confidence interval* needs to be estimated.

3.1. Examples

Table 1 gives a very simplified example of the result of a randomised trial comparing two different treatments measuring the effect of treatment dichotomously. The reason to conduct a trial would normally be the belief or hope that the new treatment has a better effect than the standard treatment. Nevertheless, one-sided tests are not recommended, and the (two-sided) hypothesis will be that the effect of two treatments differs.

Table 1
Worked example of number of responders on two different treatments ($n = 400$).

	Response	No response	Total
New treatment	156	44	200
Control group	132	68	200
Total	288	112	400

Since the hypothesis to be tested must be the opposite of what we hope to demonstrate, the *null hypothesis* (H_0) will be that the two treatments do not differ. Testing this null hypothesis with a chi-square test leads to $p = 0.008$. In other words, it is highly unlikely to see such a large difference in the proportion of responders just by chance. Since $p < 0.05$ H_0 is rejected.

An obvious question is whether the observed difference is large enough to be of *clinical relevance*. The proportion of responders in the group receiving the new treatment is 156/200 = 0.78, compared to a response rate of 132/200 = 0.66 in the control group. The conclusion with regard to clinical benefit of a 12% absolute difference would depend on the actual clinical setting and perhaps the uncertainty with which this difference is estimated. The corresponding 95% confidence interval is [0.03 to 0.21], so the true difference might indeed be quite small and close to 0. Importantly, the confidence interval does not cover the value 0, so the true effect is not likely to be 0. This corresponds to the result of the significance test ($p < 0.05$) where the null hypothesis of no difference was rejected.

3.2. Effect estimate

The *p*-value is closely linked to sample size. This can be illustrated by reducing the sample size while keeping the proportion of responders constant. In Table 2, the proportions are 0.78 and 0.66, as before, but the sample size is ¼ of that in Table 1. With exactly the same difference in effect, the *p*-value now becomes $p = 0.18$ and the null hypothesis cannot be rejected. Correspondingly, the 95% confidence interval is now estimated to [−0.06 to 0.29] and a 0 difference cannot be excluded. This neatly illustrates that no conclusion should be drawn on the basis of a *p*-value only. A small trial will rarely detect even large effects of obvious clinical relevance. On the other hand inclusion of a very large number of patients may easily lead to small *p*-values even if the difference in effect is of no or marginal relevance. In order to assess whether a statistically significant finding is of clinical relevance an estimate of the effect and a 95% confidence interval for the true effect is therefore needed. The width of the confidence interval is linked to sample size and variability. If the confidence interval is wide, the effect is not precisely estimated. To obtain higher precision, the sample size must be increased.

Table 2
Worked example of number of responders on two different treatments ($n = 100$).

	Response	No response	Total
New treatment	39	11	50
Control group	33	17	50
Total	72	28	100

Table 3
Examples of number of patients needed to detect differences.

Proportion of responders		Clinically relevant difference	Total number of patients	
Control group	New treatment		Power 80%	Power 90%
0.40	0.70	0.30	64	86
0.40	0.60	0.20	142	190
0.40	0.50	0.10	576	770
0.40	0.45	0.05	2164	2896

3.3. *Sample size*

The examples in Tables 1 and 2 also illustrate the need for sample size estimation before initiation of a trial. A trial should be large enough to detect effects considered to be of clinical relevance, *i.e.* if there is a true difference in effect, the probability that the trial leads to $p < 0.05$ should be high. The larger the number of patients, the higher is the power of the trial. Usually trials are planned with 80% or 90% power to detect a predefined difference. Table 3 shows the number of patients to be included for some examples of effect differences in a parallel group trial with a binomial (yes/no) response variable. It is clearly demonstrated that trials need to include a large number of patients in order to obtain an acceptable probability of detecting small differences, which might be of clinical relevance. Consequently, it is very important to preplan sample size and to ensure that a trial is large enough to lead to a valid conclusion. Small trials should be avoided, primarily because the probability of drawing a valid conclusion is low. Also, there will be low precision in the estimate of the treatment effect. On top of that, small trials increase the risk of publication bias whereby trials with significant results are more likely to be published than inconclusive ones. That is not the way to bring science forward.

3.4. *Randomisation and blinding*

An important assumption for *p*-values to be directly interpretable is that a trial is randomised, *i.e.* that patients are randomly allocated to treatment groups. In order to avoid bias the trial should also be double blind if at all possible, implying that neither the patient nor the treating physician knows which treatment the patient receives. In many cases comparisons are made based on observational rather than experimental data. That calls for caution when attempting to draw conclusions about causality. As a minimum, all relevant confounding factors, which might affect prognosis must be taken into account in the statistical analysis.

3.5. *Non-inferiority*

Not all trials are performed to demonstrate superiority of one treatment over another. The prerequisite for a marketing authorisation to be granted is that the benefit of a drug is shown to outweigh the risks. Hence it is sufficient to demonstrate non-inferiority of a new medicinal product compared to one that can be considered treatment of choice. In this situation, the null hypothesis to be tested would be that the new treatment is inferior to the standard treatment, and the alternative is that it is as good or superior (one-sided hypothesis test). As identical effect can never be proven, it must be accepted that the new treatment could be slightly inferior to the standard treatment, but this difference (usually called delta – δ) must be so small that it can be regarded not to be of importance to patients. In practice decisions are made based on the 95% confidence interval for the difference in treatment effect (new – standard). If the interval includes the pre-defined value δ, the effect of the new treatment is not non-inferior. If the lower limit of the confidence interval is larger than δ, however, the new treatment is considered as non-inferior to the standard treatment and could have a marketing authorisation if the quality and safety of the medicinal product were considered acceptable.

3.6. *Multiple testing*

One of the important weaknesses of using *p*-values (or indeed confidence intervals) as a means of assessing efficacy is the risk of drawing wrong conclusions. Applying a significance level of 5% implies that the risk of wrongly concluding that there is a difference between groups is 5% *for each p-value calculated.* This is often referred to as the type I error rate. If we perform a number of significance tests, the risk that one or more will become significant just by chance increases with the number of tests performed. The most common way of avoiding inflation of the probability of a type I error is to define one primary endpoint or outcome variable of a trial and to restrict the number of secondary endpoints and to regard these as supportive evidence only. That implies that if no difference in effect can be demonstrated on the primary endpoint, one should be extremely cautious to claim efficacy based on a statistically significant finding on one of several secondary variables. A very simple way to adjust for so called multiple tests is to apply a Bonferroni correction, *i.e.* multiplying all p-values with the number of tests performed. This reduces the risk of a false positive conclusion, but at the same time reduces power.

Multiplicity problems do not only arise with assessment of multiple endpoints or outcome variables. Also subgroup analyses, based on biomarkers or other patient characteristics, pairwise comparisons of more than two groups, as well as the use of different cut-offs of a continuous variable to minimise the *p*-value, will lead to inflation of the type I error rate. So will interim analyses and analyses of repeated measurements over time if an appropriate analysis method is not applied. Mills [1] correctly stated “if you torture your data long enough, they will tell you whatever you want to hear”. The best advice is to regard any *post hoc* or data driven analyses as exploratory rather than confirmatory. Since even rare events may occur by chance, unexpected findings need to be tested in independent data sets or future trials.

3.7. *Missing values*

The majority of clinical studies will not succeed in having all information on all variables registered for all patients and a plan for the handling of missing values is needed. In principle there are two different strategies that can be followed. One is to analyse complete data only, often referred to as per-protocol analysis in the clinical trial setting. At first sight, that seems to be practical, but such a strategy carries a large risk of bias. In order to obtain an unbiased estimate of effect, all missing values must be plausibly assumed to be missing at random which is rarely the case. If patients withdraw from a clinical trial due to lack of efficacy but are not included in the statistical analysis, the efficacy estimate could obviously be biased.

3.7.1. *Intention-to-treat analysis*

The recommended strategy is therefore a so-called *intention-to-treat analysis* whereby all patients are included in the analysis whether or not they received treatment as planned and have all variables of interest registered (see for instance the ICH E9 guideline [2]). Such an analysis strategy requires imputation of missing values by making assumptions on what the plausible value would be. A commonly used method is carrying the last registered value forward (LOCF). Others include carrying the baseline value forward (BOCF) or imputing the worst possible value. More sophisticated methods like multiple imputation techniques usually improve the results. In any case, all imputation strategies have weaknesses and a set of sensitivity analyses applying different methods should be performed to demonstrate that the result is robust and does not depend heavily on the way missing values are handled.

4. Summary and recommendations for clinical scientists

Some very important questions need to be answered before initiating a clinical trial. What is the research question? To which patients should the result be generalised? Is the number of patients sufficient to draw a valid conclusion? When data are analysed the number of (preplanned) significance tests should be kept small and *post hoc* analyses should be avoided. It should also be remembered that the clinical relevance of a finding cannot be assessed by the *p*-value. Thus effect estimates and corresponding 95% confidence intervals should always be reported.

Conflict of interest

No conflict of interests declared.

References

[1] Mills JL. Data torturing. N Engl J Med 1993;329:1196–9.

[2] ICH E9. Note for guidance on statistical principles for clinical trials (CPMP/ICH/363/96); 2013. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf