Australian Critical Care (2010) **23**, 93—97

ELSEVIER

Australian College of Critical Care Nurses

www.elsevier.com/locate/aucc

STATISTICS PAPER

Statistical and clinical significance, and how to use confidence intervals to help interpret both

Judith Fethney B.A. (Hons)*

Sydney School of Nursing, University of Sydney, Sydney, Australia

Received 21 May 2009; received in revised form 7 February 2010; accepted 2 March 2010

KEYWORDS
Statistical;
Clinical;
Significance;
Confidence intervals

Summary Statistical significance is a statement about the likelihood of findings being due to chance. Classical significance testing, with its reliance on *p* values, can only provide a dichotomous result — statistically significant, or not. Limiting interpretation of research results to *p* values means that researchers may either overestimate or underestimate the meaning of their results. Very often the aim of clinical research is to trial an intervention with the intention that results based on a sample will generalise to the wider population. The *p* value on its own provides no information about the overall importance or meaning of the results to clinical practice, nor do they provide information as to what might happen in the future, or in the general population. Clinical significance is a decision based on the practical value or relevance of a particular treatment, and this may or may not involve statistical significance as an initial criterion. Confidence intervals are one way for researchers to help decide if a particular statistical result (whether significant or not) may be of relevance in practice.

Introduction

Not everything that can be counted counts, and not everything that counts can be counted (attributed to Einstein).

While Einstein may have been pondering on the anomalies of the universe, the quote has relevance for the distinction between statistical significance and clinical significance. In many people's minds, the word 'significant' means 'important', but in the world of statistics, it is a statement about the likelihood of a result being due to chance, or the amount of uncertainty we are prepared to accept, not its importance. In the world of clinical practice, whether or not a result is significant is based on its importance to, and implications for, practice; that is, the practical value of any particular result.

Over the past few decades there has been considerable discussion relating to the interpretation and limitations of significance testing.[1—3] In recognition that classical significance testing cannot answer

* Tel.: +61 2 9351 0738; fax: +61 2 93510779.
E-mail address: jfethney@sydney.edu.au.

doi:10.1016/j.aucc.2010.03.001

the '*so what*' question, there have been calls for researchers to demonstrate the practical, or clinical significance of their findings. Many researchers now provide, in addition to either *p* values or alpha levels (and sometimes both), measures of effect size and confidence intervals, which provide additional information as to whether any particular *p* value returned by a test may have relevance to practice. Other researchers, in the absence of any statistically significant results, may still claim clinical significance based on a judgement about the relevance of a finding to his or her practice.

Statistical significance

Open almost any scientific journal reporting original empirical research, and we are likely to find a statement such as 'An alpha level of <0.05 was selected as the maximum value for statistical significance', or 'a statistically significant *p* value of <0.001 was found'. According to the literature, there is considerable general confusion between *p* values and alpha levels both in their application and interpretation.[4–6] There has also been a longstanding debate as to the overall role and utility of classical significance testing.[7,8] This article will not weigh into the debate surrounding the confusion, or the superiority of one approach over the other, or whether there is still a place for significance testing (the interested reader is directed to the above references). Rather the aim here is to report what researchers usually do and to demonstrate that classical significance testing does not necessarily allow researchers to make decisions about the clinical significance of their results.

An alpha level is a decision rule, specified in advance of any statistical testing, about accepting or rejecting a null hypothesis when the null hypothesis is 'true' for the population. Alpha levels are a statement about the risk we are prepared to accept in making an error in either accepting or rejecting the null hypothesis.[9] By convention and quite arbitrarily, most researchers consider an alpha level of 0.05 to be the maximum value for statistical significance, although a researcher may specify an alpha of 0.01, or 0.001, depending on how critical the results are deemed to be. If the actual *p* value returned by a statistical test is less than the alpha level, the null hypothesis is rejected; if the *p* values are greater, the null hypothesis is accepted. If framed within a null hypothesis, all *p* values less than the alpha level are as equally significant, as all cause the null to be rejected. With alpha specified as 0.05, whether a *p* value comes back as equal to 0.043, 0.027, or 0.002, they are all as equally significant as each other; it cannot be said that $p = 0.002$ is 'more significant' than $p = 0.027$.

Not all researchers specify a null hypothesis. A researcher may just ask a research question, collect and analyse the data, and see what *p* value is returned by the relevant statistical test, although they are probably bearing some alpha level in mind. A *p* value is the probability of finding a result as extreme, or fantastic, or disappointing, as the one returned by a statistical test. Consider the results of hypothetical research into the effectiveness of two hypertensive agents. This research was not framed in terms of a null hypothesis; rather a research question was asked, such as 'Which drug is more effective in reducing arterial blood pressure, A or B?

Results showed that Group A, who received drug A, had arterial blood pressure that was 3 mm Hg lower than those in Group B who received drug B. Statistical testing showed the differences in arterial blood pressure to be statistically significant at p <0.018. This means that the probability of obtaining this result was <0.018, or, put another way, <1.8%. As this is a low probability the researchers concluded that chance is unlikely, and that the difference in scores is likely to be due to the administration of drug A. However, a difference of 3 mm Hg is trivial and was not found to equate to any improved outcomes for patients. The result may have been statistically significant, but clinically non-significant. We are left asking '*so what*?', and the *p* value cannot help us answer that question. There may be a number of reasons why the difference in arterial blood pressure was significant, large sample size being one of them. The difference in blood pressure, although statistically significant, just did not count.

Clinical significance

In 1984 Jacobsen et al.[10] first proposed clinical significance as a way to determine the practical value of a treatment, as opposed to the statistical significance. According to LeFort[11] clinical significance should reflect 'the extent of change, whether the change makes a real difference to subject lives, how long the effects last, consumer acceptability, cost-effectiveness and ease of implementation'. Clinicians are often more interested in these aspects than whether the observed result was likely to be due to chance.

Although well established conventions for demonstrating statistical significance exist there are no such guidelines for the quantification of clinical significance. Some authors argue that findings cannot be clinically significant if they have occurred by chance, making statistical significance a necessary condition for the determination of clinical significance.[12] Others have found statistical significance does not necessarily equate to clinical significance.[13,14] Irrespective of whether statistical significance was proposed as a requirement for clinical significance or not, all the above authors emphasised that for there to be an assessment of the practical value of results there must be a definition of clinical significance established for the particular outcome measure used.

One way in which researchers establish clinical significance is to determine the minimum important difference (MID). This is the smallest difference in scores between groups on a particular outcome measure that would be of interest.[15,16] Three ways in which the MID for an outcome measure can be determined are to use anchor-based, distribution-based or expert panel approaches. Anchor-based approaches compare the change in the outcome of interest to some other measure of change, considered an anchor. There must be a measure of association between the outcome of interest and the anchor.[17] For example, Farrar et al.[18] compared patients' subjective ratings on several pain scales with the amount of administered medication, an objective anchor, to determine a clinically significant improvement in self-reported pain.

Distribution-based approaches are based on the statistical properties of the scale used to measure the outcome, such as the effect size, standard deviation or standard error of measurement.[19] After conducting a systematic review of a range of health-related quality of life measures Norman et al.[20] concluded that a difference of half a standard deviation represented the MID. Anchor and distribution-based approaches, the variety of methods available within them and their limitations are discussed in Copay et al.[21]

The expert panel approach invites experts in the field to read relevant literature and attempt to reach consensus as to the MID.[22] It is recommended that all researchers consider the clinical significance of their research and how this might be determined. What particular approach to use, or combination of approaches, will be highly dependent on the objectives of the research, the target group, the instruments used, whether data are scored as continuous or categorical, the normality of the data and the researchers themselves.

Confidence intervals

Very often, researchers want to generalise their results to the wider population; to change a particular practice, we have to know whether or not the new practice will, in some way, be of greater benefit than the old practice. We can estimate sample means, or the differences between two or more sample means, or a correlation, or the odds ratio of an event occurring within the sample, but the important point is that they are relevant to a specific sample only — these estimates do not tell us what the actual values might be in the wider population from which the sample is drawn. One of the limitations of the *p* value is that it is a sample estimate only.

Confidence intervals are one way to help a researcher assess what the values might be in the wider population. Confidence intervals provide the plausible range of values, bracketed by lower and upper limits that encompass the unknown population or 'true' value estimated by that sample mean, correlation coefficient or odds ratio. It is usual to report either the 90%, 95% or 99% confidence interval (CI); the 95% CI tends to be the one commonly used. If we estimate a 95% CI, then we are saying there is a 95% probability that the interval, bounded by the lower and upper limit, contains the 'true' population value, or parameter, with a 5% probability that the interval does not contain the population value. Our sample estimate will be right in the middle of the confidence interval.

Confidence intervals can be used descriptively. Consider hypothetical results from retrospective research conducted in a major metropolitan hospital on adverse events affecting patients in ICU. Per 100 patient days there was a mean of 28 adverse events, with a standard deviation of 4.6 and standard error of 1.57. The standard error is the sample standard deviation divided by the square root of the number of observations in the study, or $\text{SE} = \text{std}/\sqrt{n}$. The standard error is an approximation of what the standard deviation of the *population* mean would be if we could sample the entire population (which we usually cannot do, so we use an approximation that takes into account the sample standard deviation and the sample size). The calculation of confidence intervals is based on the assumption that the distribution of the variable being measured approximates a normal curve. In a normal curve, 68% of all observations lie within ± 1 standard deviation from the mean, 95% of all observations lie within ± 1.96 standard deviations, and 99% of observations lie with ± 2.58 standard deviations. When calculating the 95% CI, 1.96 is used as the multiplier.

Box 1

Method: Premature infants were randomly assigned to one of two treatment arms within four separate weight gain interventions. Infants could be in one trial only. Within each intervention, group A received the intervention, and Group B received standard treatment. Weights were compared using the Mann—Whitney *U*-test. In each intervention a weight gain difference of 500 g or more was considered to be clinically significant. A negative lower limit implies a weight loss.

Results: The *p* values, mean difference in weight gain and associated 95% CIs are show below.

Weight gain intervention	*p* value	Mean difference in weight gain between the two groups in each intervention (g)	95% CI for the mean difference in weight gain between the two groups in each intervention (g)
1	<0.05	885	490—1280
2	<0.05	200	180—220
3	>0.05	190	—100—300
4	>0.05	510	0—1020

To calculate the 95% CI for the mean number of adverse events described above, the standard error of 1.57 is multiplied by 1.96, which equals 3.08. This value is then subtracted from the mean to give the lower limit, and added to the mean to give the upper limit. The 95% CI for the mean number of 28 adverse events is therefore 24.92—31.08, indicating there is 95% probability that the mean number of adverse events in ICUs in the wider population lies between 24.92 and 31.08. This may be important information for the planning of care, for professional development of staff, for comparison with ICU units in other hospitals nationally and internationally or for changing practice.

How statistical significance, clinical significance and confidence intervals can work together

More typically, confidence intervals play a part in statistical testing, and they can also be used to determine clinical significance. The results of a hypothetical randomised control trial are shown in Box 1. In this example, premature infants in neonatal intensive care were randomly allocated to one of four weight gain interventions. Within each intervention, infants were allocated to either receive the proposed intervention, or standard care. This was a large, international study and the decision to consider as clinically significant a 500 g difference in weight gain over a three-month period between premature babies who received the intervention and premature babies who did not, was made by referring to previous published research and an expert panel of neonatal clinicians. The null hypothesis in all instances was that, within each of the four specific interventions, there would be no difference in weight gain between infants who received the intervention and those who received standard care.

To be statistically significant, the CI must not include zero, as this would indicate no difference in weight gain between the two groups being compared within each intervention. To be considered for clinical significance, the lower limit of the CI must be equal to or greater than 500. Interventions 1 and 2 were both statistically significant, as the difference in weight gain between the two groups within these interventions did not include zero. The CI for intervention 1 shows a 95% probability that in the wider population premature infants who receive this intervention are likely to gain between 490 and 1280 g more than infants receiving standard treatment, therefore within the realm of clinical significance. For intervention 2, however, there is a 95% probability that the mean weight gain difference would be between 180 and 220 g, clinically non-significant. Interventions 3 and 4 were statistically non-significant, as the CIs for both included zero (no difference). The CI for intervention 3 shows that the weight gain difference in the wider population would not be clinically significant (in fact, some babies could lose weight), while intervention 4 shows the weight gain difference *could* be clinically significant, as some infants in this intervention did gain more than the required 500 g difference. Research can then focus on those infants in intervention 4 who achieved the required weight gain, and use this information to improve and/or better target post natal care for premature infants.

Aspects such as sample size are likely to have an effect on the *p* values. Intervention 2 may have had a large number of participants, so that even relatively trivial weight gain was detected as significant, while intervention 4 may have had a sample size that was too small. Provision of the confidence intervals allows an assessment of the practical significance of the results. To halt interpretation at the *p* values affords them more meaning than they deserve, and we are at risk of assuming intervention 2 to be more effective than it was, and to discount intervention 4 when it may actually be of considerable benefit to some premature infants. Even though the *p* value says the result is statistically non-significant, it still might 'count' for some infants. We can see clearly that more information over and above the *p* values is required to base sound decisions on either statistical, or clinical significance, or both.

Conclusion

Reliance on *p* values only allows a dichotomous decision — statistically significant or not. While sometimes a yes/no decision may be the research objective, such thinking can hinder clinicians from reflecting and appreciating what their data really mean. Identifying clinical importance is what clinicians are ultimately aiming for, not the identification of statistical significance. Researchers should always determine the minimum important difference of the outcome under study and present confidence intervals where possible as they enable the identification of potential clinical significance, even in the absence of statistical significance.

References

1. Cohen J. The earth is round ($p < .05$). *American Psychologist* 1994;**49**:997—1003.
2. Loftus GR. Psychology will be a much better science when we change the way we analyse data. *Psychological Science* 1996;**7**:161—71.
3. Hubbard R, Lindsay RM. Why *p* values are not a useful measure of evidence in statistical significance testing. *Theory Psychology* 2008;**18**(1):69—88.
4. Hubbard R, Ryan PA. The historical growth of statistical significance testing in psychology—and its future prospects. *Educational and Psychological Measurement* 2000;**60**:661—81.
5. Hubbard R. Alphabet soup: blurring the distinctions between *p*'s and α's in psychological research. *Theory & Psychology* 2004;**14**(3):295—327.
6. Christensen R. Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician* 2005;**59**(2):121—6.
7. Harlow LL, Mulaik SA, Steiger JH, editors. *What if there were no significance tests?*. Mahwah, NJ: Erlbaum; 1997.
8. Schmidt FL. Statistical significance testing and cumulative knowledge in psychology: implications for the training of researchers. *Psychological Methods* 1996;**1**:115—29.
9. Pereira S, Leslie G. Hypothesis testing. *Australian Critical Care* 2009;**22**(4):187—91.
10. Jacobsen NS, Follette WC, Revenstorf D. Psychotherapy outcome research: methods for reporting variability and evaluating clinical significance. *Behavior Therapy* 1984;**15**(4):336—52.
11. LeFort S. The statistical versus clinical significance debate. *Journal of Nursing Scholarship* 1993;**25**(1):58.
12. Greenstein G. Clinical versus statistical significance as they relate to the efficacy of periodontal therapy. *Journal of the American Dental Association* 2003;**134**:583—91.
13. Todd KH, Funk KG, Funk JP, Bonacci R. Clinical significance of reported pain severity. *Annals of Emergency Medicine* 1996;**27**(4):485—9.
14. Sackett DL. Superiority, equivalence and noninferiority trials. In: Haynes RB, Sackett DL, Guyatt GH, Tugwell P, editors. *The principles behind the tactics of performing therapeutic trials. Clinical epidemiology: how to do clinical practice research*. 3rd edition Philadelphia, PA: Lippincott Williams & Wilkins; 2006. p. 196—206.
15. Redelmeier DA, Guyat RA, Goldstein DS. Assessing the minimal important difference in symptoms: a comparison of two techniques. *Journal of Clinical Epidemiology* 1996;**49**(11):1215—9.
16. Wright JG. The minimal important difference: who's to say what is important? *Journal of Clinical Epidemiology* 1996;**49**(11):1221—2.
17. Busse JW, Guyatt, GH. Optimizing the use of patient data to improve outcomes for patients: narcotics for chronic noncancer pain. *Expert Review of Pharmacoeconomics Outcomes Research* 2009;9(2):171—9. Available at http://www.medscape.com/viewarticle/705606 [Accessed 4 February 2010].
18. Farrar JT, Portenoy RK, Berlin JA, Kinman JL, Strom BL. Defining the clinically important difference in pain outcome measures. *Pain* 2000;**88**:287—94.
19. Yost KJ, Eton DT. Combining distribution and anchor-based approaches to determine minimally important differences: the FACIT experience. *Evaluation and the Health Professions* 2005;**28**:172—91.
20. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Medical Care* 2005;**41**(5):582—92.
21. Copay AG, Subach BR, Glassman SD, Polly DW, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. *The Spine Journal* 2007;**7**:541—6.
22. Ostelo R, Deyo RA, Stratford P, Waddell G, Croft P, Von Korff M, Bouter LM, de Vet HC. Interpreting change scores for pain and functional status in low back pain. *Spine* 2008;**33**(1):90—4.