Clin Orthop Relat Res (2009) 467:2736–2737
DOI 10.1007/s11999-009-0952-1

EDITORIAL

Reporting Results of Orthopaedic Research

Confidence Intervals and p Values

Raphaël Porcher PhD

Published online: 30 June 2009

The paper of Vavken et al. published in this issue of *Clinical Orthopaedics and Related Research* [7] underscores the low frequency of reporting confidence intervals in orthopaedic research as opposed to reporting p values, despite recommendations in the medical literature to report the former [3–6]. One possible reason for the investigators favoring statistical testing (ie, p values) over confidence intervals may be the lack of understanding of the usefulness of confidence intervals, and the habit of seeing p values in almost all surgical papers. Although reporting confidence intervals frequently is advisable, doing so should not be in opposition to p values, as both correspond to different aims, namely, estimation and hypothesis testing, and both convey different although related information.

Suppose a randomized trial comparing the postoperative Harris hip score at 3 months between two groups of 100 patients undergoing two different surgical procedures, where the observed mean Harris hip scores are 80 and 90 in each group, respectively, and the standard deviations 20 in both groups. To make a judgment regarding the superiority of one procedure over the other, we would need the true or population difference between both, ie, the difference that would have been obtained if all eligible patients had been included in the study rather than just a sample of them [3]. The trial only yields an estimate of this population difference, which here is equal to 10. Such a value alone, however, is of little help in determining if one surgical procedure is superior to the other, because we have no information on where the true difference lies, whether this difference of 10 could have been observed by chance only, and whether real differences are clinically meaningful. The first question relates to the confidence interval, the second to the p value, and the third to clinical judgment based on other information.

The confidence interval of a parameter, such as the difference in means of our example, is that range of values in which we are confident that the true or population value of the parameter lies. The level of confidence is chosen by the investigator, usually at 95%, although values of 90% or 99% sometimes are used. In our example the 95% confidence interval of the difference in mean Harris hip scores would be 4.4 to 15.6. A 90% confidence interval would be 5.3 to 14.7, which is narrower than the 95% confidence interval; greater confidence being obtained by wider intervals. Details regarding how to compute these intervals were presented by Gardner and Altman [3], but they depend on the observed mean difference between groups, the standard deviation (or equivalently the variance) of the Harris hip score, the sample size in each group, and the confidence level. In particular, a confidence interval will be narrower for larger sample sizes [2]. The result of the trial is thus that we can be 95% confident that the true difference in means scores obtained by each procedure lies within 4.4 to 15.6. From a statistical point of view a 95% confidence interval means that if we had repeated the trial in the same population a very large number of times and computed a 95% confidence interval for each trial, then 95% of these confidence intervals would include the true difference between the mean scores.

The author certifies that he or she has no commercial associations (eg, consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article.

R. Porcher (✉)
Département de Biostatistique et Informatique Médicale, Hôpital Saint-Louis, Inserm UMR-S 717, Université Paris Diderot, 1, avenue Claude Vellefaux, 75475 Paris Cedex 10, France
e-mail: raphael.porcher@univ-paris-diderot.fr;
raphael.porcher@paris7.jussieu.fr

Using a t test to test the null hypothesis that the true difference between mean scores is zero would yield a p value of 0.0005. This means that the difference of 10 between groups we observed or a larger difference only had a (very low) probability of 0.0005 if the null hypothesis of no difference were true [2]. Note the difference with a common misinterpretation of p values in terms of probability of no difference between groups. However the lower the p value, the more unlikely the null hypothesis is. As the p value is less than 0.05, the difference is said to be significant at the 5% level, and the null hypothesis of no true difference is rejected at such a level. Similarly to the confidence interval, the p value depends here on the observed difference, the standard deviation, and the size of each group.

The confidence interval and a hypothesis test are linked: both depend on the same quantities, and the result of the hypothesis test at a given level can be inferred from the corresponding confidence interval. In our example, the zero difference between means corresponding to the null hypothesis is outside the 95% confidence interval, which indicates that the t test will reject the null hypothesis at the $(100 - 95) = 5\%$ level.

Conversely, reporting a p value alone does not provide much additional information. The confidence interval shows the range of true values of the parameter compatible with the study results on the same scale as the end point analyzed and allows direct interpretation of the magnitude of the effect, whereas the apparent precision of a p value (eg, one could obtain $p = 0.00153$) does not allow us to judge the clinical relevance of the effect. In our example, with a much larger sample size (1000 patients per group), a mean difference of 2 could have led to a 95% confidence interval of 0.2 to 3.8. Despite a significant hypothesis test at the 5% level, the confidence interval also would have shown that the difference between the two surgical procedures was unlikely to be clinically important. An even more frequent problem arises with misinterpreting a nonsignificant hypothesis test [1]. In many cases, researchers have concluded at a similar effectiveness or at no relationship between two variables as soon as no statistically significant effect was found, even when confidence intervals would have included clinically meaningful differences or associations. For instance, had the observed difference between the two groups of our example been 5, the 95% confidence interval would have been -0.6 to 10.6, and the p value 0.079. We thus would have concluded a nonsignificant difference. However, if a real difference of 10 between both surgical procedures is considered clinically meaningful, then the study does not rule out such a difference. Of course, one should not presume a true effect from a large observed difference if the associated test is not significant. Rather, an insignificant test with a confidence interval comprising clinically relevant differences suggests the study had insufficient power to detect effects [2] and the information from a confidence interval therefore is crucial for correct interpretation. In all cases, a confidence interval conveys more useful information, and this is the reason why many authors advocate reporting confidence intervals. In that respect, the policy of the *British Medical Journal* to encourage the use of confidence intervals without prohibiting p values [3, 4] seems a reasonable one, and we could recommend reporting both for the major findings of a study.

Cases in which a p value would be preferable to a confidence interval are difficult to find. When more than two groups are compared, however, a global p value for the test of the null hypothesis of equality of a parameter (eg, mean or proportion) in all groups can be obtained without raising multiplicity issues (ie, the possibility of finding a statistical difference by chance when performing multiple comparisons). Computing confidence intervals for multiple pairings of a larger number of groups would need adjusting of the individual confidence levels to control a global confidence level. Nevertheless these confidence intervals still remain of interest. Multiplicity issues also arise when many variables are compared in the same study. However in this case, the investigator is confronted with the same problem of multiple comparisons whether p values or confidence intervals are used.

Finally, confidence intervals are not always appropriate, as when using descriptive statistics. As with hypothesis testing, they are part of statistical inference, ie, using the observed data to convey information on the population from which the study patients were sampled. For example, confidence intervals thus do not adequately describe how patient values are distributed. Confidence intervals and p values also only account for the effects of sampling variation on the precision of the estimated parameter but cannot control for biases in sampling or study conduct.

References

1. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ*. 1995;311:485.
2. Biau DJ, Kernéis S, Porcher R. Statistics in brief: the importance of sample size in the planning and interpretation of medical research. *Clin Orthop Relat Res*. 2008;466:2282–2288.
3. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)*. 1986;292:746–750.
4. Langman MJ. Towards estimation and confidence intervals. *Br Med J (Clin Res Ed)*. 1986;292:716.
5. Rothman KJ. A show of confidence. *N Engl J Med*. 1978;299: 1362–1363.
6. Thompson WD. On the comparison of effects. *Am J Public Health*. 1987;77:491–492.
7. Vavken P, Heinrich KM, Koppelhuber C, Rois S, Dorotka R. The use of confidence intervals in reporting orthopaedic research findings. *Clin Orthop Relat Res*. 2009. doi:10.1007/s11999-009-0817-7 (this issue).