Editorial

P-values in research reports

Reports from empirical medical research can, generally speaking, be grouped into one of two categories: the case report and the analytical study report. The former is typically a descriptive report presenting observations on a single patient together with the author's comments. The latter is based on observations from groups of patients and the conclusions made rest on hypothesis testing or parameter estimation.

From a historical point of view, case reports have dominated medical journals for centuries. Today, however, the analytical report is standard. This change has occurred recently—mainly during the last 40–50 years.

The education and training of medical researchers usually includes courses on statistics. Unfortunately, the focus is often on mechanical calculation of p-values rather than on understanding the fundamental principles. One consequence of this is that many medical reports present p-values by routine, with little or no consideration of the rationale behind it.

For example, p-values do not change observed data; significance tests are tools for making inferences from the studied population to an unobserved greater population of subjects which the results are being generalized to. A difference that exists in observed data thus exists whether it is statistically significant or not. The common phrase "no difference was observed" when presenting statistically insignificant, but clearly observable, differences is not a language problem: it suggests confusion of fundamental statistical principles.

Another example is the common bad habit of presenting p-values from comparisons of baseline characteristics in randomized trials. This testing is actually equivalent to testing whether randomization has taken place. With randomization and a 5% level of statistical significance, about 5% of the tests performed can, by definition, be expected to show statistical significance. A substantially higher frequency of statistically significant tests could indicate that randomization did not take place. However, the rationality of testing whether a randomized trial is randomized is indeed not obvious and should be thoroughly explained if performed.

This issue of Acta Orthopaedica contains an article (Bhandari et al. 2005), which suggests that p-values distort readers' perceptions of observed results, that statistical significance is generally mistaken for clinical significance. This can perhaps explain the phenomena described in the two preceding examples. The authors conclude that the use of p-values impairs understanding of research results, and they question the use of p-values in future: should they be abolished? Actually, some medical journals have already attempted to ban p-values (Rothman 1998, Thomason et al. 2004); confidence intervals have been proposed (Gardner and Altman 1986) as a better alternative. The results of these attempts have, however, been disappointing (Thomason et al. 2004).

However, these p-value problems should not be discussed in isolation. Performing and reporting analytical studies as if they had been case reports is common, but counterproductive. Case reports may present scientifically important observations but, in contrast to analytical studies, their primary purpose is not to generalize results beyond the observed data. The inferential aspects of analytical studies should be emphasized, not ignored.

I believe that p-values play an important role in medical research and will continue to do so in the future. However, the misunderstandings and misuses of p-values should be abandoned, and authors, reviewers and editors have a common responsibility to contribute to a better practice.

It should be appreciated that confirmatory studies generally have higher levels of evidence than exploratory ones, which are performed for the purpose of generating hypotheses, usually with less rigorous adherence to statistical precision and

Copyright © Taylor & Francis *2005. ISSN 1745–3674. Printed in Sweden – all rights reserved.*
DOI 10.1080/00016470510030715

validity. This difference should also be recognized when prioritizing manuscripts for publication.

Furthermore, it should be recognized that a confirmatory study can only answer a limited number of questions and they should be described in a protocol prior to performing the study. This study protocol should include pre-specified patient number calculations showing that the statistical precision of the study is sufficient, at least for one primary endpoint.

With this backing, we will produce more accurate research results; it will improve the general quality of reports and reduce much of the confusion and misunderstandings surrounding p-values.

Jonas Ranstam
jonas.ranstam@ort.lu.se

Bhandari M B, Montori V M, Schemitsch E H. The undue influence of significant p-values on the perceived importance of study results. Acta Orthopaedica 2005; 76: 291-5.

Gardner M J, Altman D G. Confidence intervals rather than P values. BMJ 1986; 292:746-50.

Rothman, K. Writing for Epidemiology. Epidemiology 1998; 9: 333-7.

Thomason F F, Cumming G, Finch S, and Leeman J. Editors can lead researchers to confidence intervals, but can't make them think: statistical lessons from medicine. Psychol Sci 2004; 15: 119-26.