CORRESPONDENCE

Judging a Plethora of p-Values: How to Contend With the Problem of Multiple Testing – Part 10 of a Series on Evaluation of Scientific Publications

by Dr. rer. physiol. Anja Victor, Dipl.-Stat. Amelie Elsäßer,
Prof. Dr. rer. nat. Gerhard Hommel, Prof. Dr. rer. nat. Maria Blettner
in volume 4/2010

Variables Should Be Pre-Set

I am always interested in the level of evidence of scientific medical studies—simply because it is the level of evidence that will determine my actions as a doctor, in continuing medical education, and as an academic teacher. The level of evidence of the arguments presented is also of central importance in guideline development.

The level of evidence is determined by the probability theory, more precisely the p-value, which describes the probability that different results than the reported ones may be achieved. In order to acquire confirmative level of evidence, one or more equivalent or hierarchically ordered hypotheses that are to be tested have to be set out a priori and in writing. Further, the required number of cases needs to be calculated, if necessary by taking into account multiple testing.

In addition to the results for the primary outcome variables, for which confirmative level of evidence can be achieved if they were planned as described, it is quite common for results of secondary outcome variables to be reported, which have merely hypothesis generating power. Adjusting the p-values of the results of secondary outcome variables is possible only if the variables were fixed a priori. If the secondary outcome variables are defined only at the end of the study or even after the evaluation of the collected data, the suspicion arises (and can't be cast aside easily) that the reported secondary outcome variables were selected from a plethora of possible variables, on the basis of whether they suited results that were arrived at by chance. Adjusting the p-values is not possible in that scenario.

DOI: 10.3238/arztebl.2010.0417a

REFERENCES

1. Victor A, Elsäßer A, Hommel G, Blettner M: Judging a plethora of p-values: How to contend with the problem of multiple testing — Part 10 of a series on evaluation of scientific publications [Wie bewertet man die p-Wert-Flut? Hinweise zum Umgang mit dem multiplen Testen – Teil 10 der Serie zur Bewertung wissenschaftlicher Publikationen]. Dtsch Arztebl Int 2010; 107: 50–6.

Prof. Dr. med. Dipl.-Chem. Frank Pohlandt
Leitlinien-Beauftragter der GNPJ
Fünf-Bäume-Weg 138/1
89081 Ulm, Germany
frank.pohlandt@uni-ulm.de

Differences Should Be Clarified

The authors are to be congratulated for expressing criticism of technical errors in handling significance tests. However, one aspect remained somewhat neglected: the decisive criterion for relevance is not significance but effect. How often do authors use “significant” in the sense of highly or extremely relevant, or even with star markings as for expensive brandy, and less knowledgeable readers are taken in. Sometimes it is even difficult to find the difference that was tested. It is advisable to always ask for the amount of difference in tests for differences and to relate these to the resulting mean. If the number of cases is not high enough, differences of 5% are “highly significant.” The same is true for the correlation coefficient: if $r=0.30$, 118 cases are enough to fall below the probability of an error of 0.001, but the coefficient of determination, r^2 (the effect), is only 9% in such a scenario. During the subsequent interpretation of the results it will become essential to reflect more on the relevance—and less on the significance—of that effect.

DOI: 10.3238/arztebl.2010.0417b

REFERENCES

1. Victor A, Elsäßer A, Hommel G, Blettner M: Judging a plethora of p-values: How to contend with the problem of multiple testing — Part 10 of a series on evaluation of scientific publications [Wie bewertet man die p-Wert-Flut? Hinweise zum Umgang mit dem multiplen Testen – Teil 10 der Serie zur Bewertung wissenschaftlicher Publikationen]. Dtsch Arztebl Int 2010; 107: 50–6.

Prof. Dr. med. H.-V. Ulmer
Curt-Goetz-Str. 93
55127 Mainz-Drais, Germany
ulmer@uni-mainz.de

In Reply:

We thank the commentators for their supporting letters and for indicating other important points related to multiple testing.

Unfortunately, it is not always possible to recognize selective reporting of secondary outcomes with “good” p-values. We can only appeal to the authors to be honest for the sake of research and encourage reviewers, editors and readers to ask critically whether hypotheses other than the original ones were reported and to inquire about the number of tests performed.

Quite aside from multiple testing, the problem with p-values for correlation coefficients is that the test is for deviation from a correlation coefficient of zero (null hypothesis: $r = 0$). As one commentator remarked, even a very weak correlation can give a small p-value (if the sample size is adequate). This is often ignored when interpreting the p-values of correlation coefficients.

The results of statistical tests must always be evaluated with respect to their clinical relevance. With respect to clinical relevance a p-value is not a measure, but only serves to estimate whether the probability of error lies

under a prescribed limit. The effect must accordingly be assessed as a value – such as a median, mean, relative frequency or similar. It is unfortunately the case that multiple testing is accompanied by an overestimate of effects. The overestimate is not eliminated by correcting the level or the p-values for multiple testing.

DOI: 10.3238/arztebl.2010.0417c

REFERENCES

1. Victor A, Elsässer A, Hommel G, Blettner M: Judging a plethora of p-values: How to contend with the problem of multiple testing — Part 10 of a series on evaluation of scientific publications [Wie bewertet man die p-Wert-Flut? Hinweise zum Umgang mit dem multiplen Testen – Teil 10 der Serie zur Bewertung wissenschaftlicher Publikationen]. Dtsch Arztebl Int 2010; 107: 50–6.

Dr. rer. physiol. Anja Victor

Institut für Medizinische Biometrie, Epidemiologie und Informatik
Universitätsmedizin der Johannes-Gutenberg-Universität Mainz
Obere Zahlbacher Str. 69
55131 Mainz, Germany
victor@imbei.uni-mainz.de

Conflict of interest statement

The authors of all contributions declare that no conflict of interest exists according to the guidelines of the International Committee of Medical Journal Editors.