British Journal of Ophthalmology, 1987, **71**, 645–646

Eyes or patients? Traps for the unwary in the statistical analysis of ophthalmological studies

ROBERT G NEWCOMBE[1] AND GEOFFREY R DUFF[2]

From the [1]Department of Medical Computing and Statistics, University of Wales College of Medicine, and the [2]Department of Ophthalmology, University Hospital of Wales, Cardiff

SUMMARY In reports on ophthalmological research the results of measurements on the eye are often expressed as mean and standard deviation based on m patients, n eyes (n>m). This approach leads to *t* tests that are invalid because the measurements on the two eyes of one subject are usually related, not independent. In a simulation study involving intraocular pressure data analysed in this way, the null hypothesis of no difference between groups was rejected at a nominal alpha=0·05 level in 39 out of 200 simulations; thus the true alpha was nearly 0·2. This approach is excessively prone to produce false positive results.

Scientific evaluation of ophthalmological interventions involves measurements of several types of variables. Typically, when beta blocking agents are evaluated for the treatment of raised intraocular pressure (IOP), interest centres on two types of measurement—those made on each eye such as the IOP, and systemic measurements such as heart rate and blood pressure. The danger that inappropriate statistical analysis may be carried out on the former type of data has prompted this investigation.

Material and methods

The data analysed were taken from a crossover study[1] comparing oral nadolol and topical timolol in 22 patients. Before any treatment was given the IOP was measured in all 44 eyes. For simplicity, consider random allocation of the 22 subjects to two treatment order groups of 11 patients each. There are 705 432 possible ways in which this can be done. Most of these will produce two subgroups of patients or eyes that are well matched for initial IOP. We would expect any valid test of the null hypothesis of equal mean IOP in these two subgroups of results—that is, initial comparability of these groups—to yield a statistically significant ($p<0·05$) result for only 1 in 20 of these partitions.

A simulation process was carried out as follows. The 22 subjects were allocated to two fictitious groups A and B of 11 subjects each by means of random numbers. Four ways of using an unpaired two-sample *t* test to compare the random groups were considered: (1) a comparison of IOP in 11 right eyes of patients in group A versus 11 right eyes of group B; (2) a comparison of IOP in 11 left eyes of patients in group A versus 11 left eyes of group B; (3) a comparison of mean IOP (average of two eyes) in 11 patients in group A versus 11 patients in group B; (4) a comparison of IOP in all 22 eyes of patients in group A versus 22 eyes in group B.

A nominal two-sided significance level of alpha= 0·05 was used, so that values of $|t|$ greater than 2·086 (tests 1, 2, and 3; 20 df) of 2·018 (test 4; 42 df) were judged significant.

Results

The simulation was carried out 200 times. The four methods yielded 8, 6, 9, and 39 'significant' differences, respectively. Thus the type 1 error rates produced were 0·04, 0·03, 0·045, and 0·195.

Discussion

It is apparent that, while analytical methods 1, 2, and 3 yield appropriately low frequencies of false positive results, method 4 is excessively prone to reveal an apparent difference when all we are observing is the play of chance. The simple two-sample *t* test used

Correspondence to R G Newcombe, Department of Medical Computing and Statistics, University of Wales College of Medicine, Heath Park, Cardiff CF4 4XN.

assumes statistical independence of different data values. Corresponding measurements on the two eyes of one patient will not generally yield identical values, but are far from being independent.

In an actual trial *t* tests may appear in several contexts: an unpaired test for initial comparability between groups; a paired test to assess serial changes on one treatment; and the unpaired tests comparing period differences between random groups which constitute the analytic method of choice for the crossover trial.[2] The above considerations apply equally to any of these. As a general rule any significance test in which the implied total 'sample size' exceeds the number of subjects in the study is invalid. Indeed the same considerations apply to confidence intervals, and standard deviations said to be based on m patients and n eyes (n>m) are misleading, as they represent a mixture of between-subject and within-subject variation.

It is difficult to ascertain conclusively in which existing research publications the significance tests are invalid on these grounds. In most articles we have examined the statistical methods and results have not been stated clearly enough to enable us to tell. It is the commonly accepted standard of 'proof' that is deficient. There are other essentially similar pitfalls, such as quoting and performing significance tests based on standard deviations or standard errors derived from repeated measurements on the same eyes.

What, then, should be done? In the situation covered by our simulations, while each of the approaches 1, 2, and 3 is valid, clearly method 3 is preferable, being based on all the information in the sample, and the averaging process leads to greater precision and greater power to detect a difference of a given size. It may happen that in some patients only one eye is used in the trial. The value for that eye may then be used instead of the average. In theory a weighted analysis is required, but it is likely to make little difference to the conclusions.

In our trial,[1] in which one of the treatments was systemic, there was only a modest gain in efficiency from using the two eyes in each subject. When both treatments to be compared are administered locally, and interest centres on local rather than systemic effects, there is an experimental design, the double-crossover (Duff G R, Graham P A, in preparation), which can yield greater statistical power. In its simplest form patients are randomly allocated to groups AB or BA. Group AB receives treatment A to the left eye and B to the right eye in the first period, then the reverse; group BA receives the treatment in the opposite order. A further refinement is possible if the main variable of interest, typically IOP, is measured on admission to the trial. Group AB can then receive treatment A to the poorer eye, B to the better in the first period, then the reverse, and conversely for group BA. Such designs are very efficient provided we have good reason to believe that there is no carry-over effect from our topical medication to the contralateral eye.

References

1 Duff GR, Watt AH, Graham PA. A comparison of the effects of oral nadolol and topical timolol on intraocular pressure, blood pressure, and heart rate. *Br J Ophthalmol* 1987; **71**: 698–700.
2 Hills M, Armitage P. The two period cross-over clinical trial. *Br J Clin Pharmacol* 1979; **8**: 7–20.

Accepted for publication 6 October 1986.