

Statistics in Question

SHEILA M GORE

ASSESSING METHODS— ART OF SIGNIFICANCE TESTING



Statistical tests of significance make an important contribution in finding out whether differences between treatments are genuine. The first step is calculation—some theory and careful arithmetic tell us how probable is a result as extreme or more extreme than our observations, if there is actually no difference between the treatments. The art of significance testing comes with the second step—interpreting that probability.¹⁻³ There is an EITHER/OR conclusion to a statistical argument, which is one very good reason why clinical decisions should not be made automatically on the basis of a single “statistically significant” finding—unless the significance level is very much more extreme than 0.05. Running major trials in parallel in different countries is recommended to expedite practical clinical decisions and to avoid ethical problems.

Interpreting significance tests calls for (a) good scientific judgment setting the results of this experiment in perspective: is the treatment a rational choice, are conflicting or corroborative reports of it already published, is the treatment effect accentuated in higher-risk patients; (b) awareness of what does not constitute good *prima facie* evidence: unexpected associations, benefit from treatment in an isolated subgroup only, improper repeated significance testing, and the danger that a publication bias favours positive findings; (c) complete reporting, including descriptive statistics⁴ and estimation of the confidence interval. The last is discussed in more detail in the next article. Estimating the interval ensures that non-significance is not mistaken for “no difference.” Wide confidence limits are usually the hallmark of inadequate trials.

Two-tailed tests of significance predominate in medicine because the possibility of an experimental treatment being inferior cannot reasonably be excluded at the start of a clinical trial. The test region therefore comprises large positive or

negative differences between treatments. Strong *a priori* grounds that favour one treatment indicate a one-tailed test of significance—looking for a treatment difference in a specified direction—but are also an ethical contraindication to a randomised clinical trial.

(8) Interpret $p < 0.05$ and $p < 0.01$: given identical trial size, which gives stronger evidence against the (null) hypothesis that there is no difference between treatments?

—if there is truly no difference between treatments an outcome as extreme or more extreme than that observed would occur fewer than: 5 times in 100 $p < 0.05$; 1 time in 100 $p < 0.01$

—an outcome that would occur less often than 1 time in 100 when there is actually no difference between treatments is more extreme (that is, less compatible with the hypothesis of no treatment difference) than an outcome that arises perhaps as often as 5 times in 100.

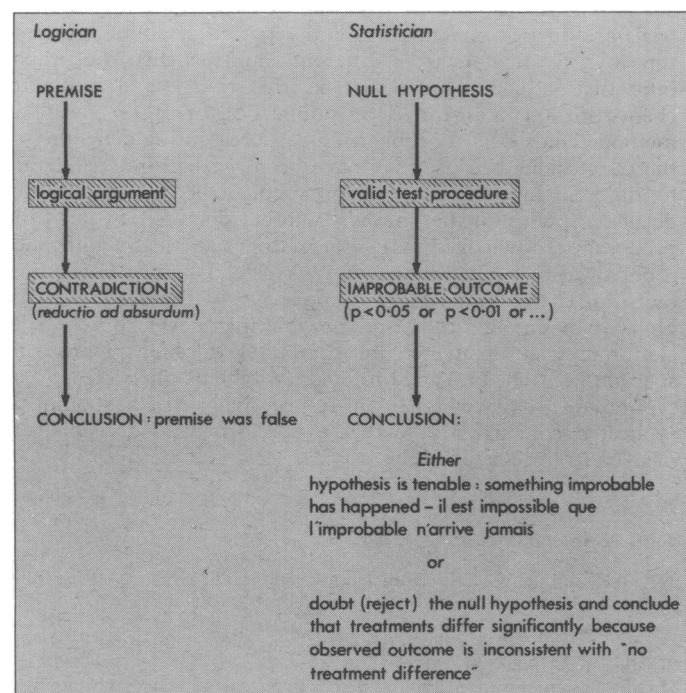
COMMENT

Proof by contradiction is a major way of tackling logical and mathematical problems. Stating as his premise what he wants to disprove the logician argues correctly from that starting point and knows that if his arguments lead him to a contradiction then, since his method was correct, the premise must have been false.

In statistics we copy this approach but instead of reaching an absolute contradiction we observe an improbable outcome. Starting from a null hypothesis—that there is no difference between treatments, for example—we observe the result of a well-designed experiment, assess how likely the observed result is *from the standpoint of no treatment difference*, and if it is judged by a valid test procedure to be an improbable outcome then *either* we accept that there is really no difference between treatments and the improbable has happened—as it must occasionally do—or we argue that because the observed outcome is unusual (improbable) if there were no difference between treatments it is, on the contrary, plausible that there *is* a difference.

How large that difference is likely to be is reflected by the (confidence) interval from the smallest to the largest effect of treatment with which the trial data are consistent. That is to say, if I took as my null hypothesis that the effect of treatment was any value in the quoted interval then my test procedure would not indicate that the result of the experiment was untoward.

The *either/or* conclusion to a statistical argument explains why clinical decisions are not made on the basis of a single “statistically significant” finding, unless the significance level is very much



more extreme than 0.05. Moreover, Zelen has warned that because authors and editors are reluctant to publish non-significant comparisons articles in medical journals may abound with false-positive claims—there are few really promising treatments and most comparisons are therefore made between

Clinical decisions are seldom made on the basis of a single "statistically significant" finding

equal contenders with only the 1 in 20 trials which fortuitously achieves $p < 0.05$ being reported. Until editors insist on confidence intervals when publishing non-significant differences^{2 5} rejoinders in the correspondence section are not the solution to Zelen's paradox. Non-significant findings from inadequate trials masquerade there as convincing counter-arguments because there is no obligation to report what are undoubtedly hopelessly

Editors should insist on confidence intervals when publishing non-significant differences

wide confidence intervals. Effective studies are undervalued and less informative than they might be because authors do not emphasise that narrow confidence limits mean precise estimation.

(9) What is evidence of a treatment effect or of association?

In addition to statistical significance on the credit side:
—treatment rationale

—dose-response relation

—effect of treatment evident in subgroups as well as in the trial as a whole

—epidemiological evidence

—more than one trial confidently pinpoints effect of treatment

—a single major trial and corroborative reports

On the debit side:

—an unexpected association needs to be checked with new data

—overall no significant treatment difference, but a significant effect in one subgroup

—eager and frequent perusal of accumulating data

Major trials should be run concurrently in different countries so that practical clinical decisions are delayed as little as possible

COMMENT

Statistical significance is necessary evidence but is not usually sufficient to change clinical practice unless one or more of the following conditions apply: (a) the treatment is a rational choice, fitting in with some theory about the disease process or because it has been shown to work for a related condition; or (b) a dose-response relation can be shown; or (c) the effect of treatment is evident in subgroups of patients as well as in the sample as a whole; or (d) epidemiological evidence can be adduced as in the report linking coffee and cancer of the pancreas⁶; or (e) more than one powerful clinical trial has been reported confidently pinpointing the treatment effect⁷; or (f) a single major trial has shown an overwhelming difference⁸ and other reports—of related treatments or similar trial end points—are corroborative; or (g) combinations of these. Extra confirmatory evidence is needed because of the *either/or* conclusion to a statistical argument.

It is important also to recognise what is not good evidence. On the debit side therefore are the following. (a) An unexpected association, discovered only on careful scrutiny of the results and not one which the investigators were originally interested in. Such an association is that between coffee drinking and cancer of the pancreas. MacMahon *et al*⁶ correctly emphasised the need for independent validation. (b) Clinical trials showing overall no significant difference between treatments but parading a significant effect in one subgroup of patients. Beware of multiple significance testing.⁹ How many subgroups were examined? This information is important because fortuitously one group out of 20 will show a treatment effect ($p < 0.05$), even when there is actually no treatment difference. Only an independent check will tell whether the treatment really is effective for this type of patient. (c) Eager and frequent perusal of accumulating data,¹⁰ the authors reporting the moment that "statistical significance" is first achieved. The chance of finding $p < 0.05$ at some time during a trial between equivalent treatments is close to 0.20,⁷ whereas if up to 10 repeated significance tests are made using

A good general principle is to view sceptically the significance level $p < 0.05$ if you suspect improper repeated significance testing

the criterion $p < 0.01$ then the effective false-positive rate is still less than 0.05. A good general principle therefore is to view sceptically the significance level $p < 0.05$ unless you can be sure that improper repeated significance testing is not responsible for it.

In relation to timolol in the treatment of patients after myocardial infarction Mitchell¹¹ asked: are consistent trends from many trials more convincing than a single trial with an extreme significance level such as $p < 0.001$? The answer is twofold. Firstly, a proliferation of inadequate trials is bad science, and frequent repetition of major trials is unethical. The second answer lies in points *a* to *g* above of what is good evidence

Proliferation of inadequate trials is bad science; frequent repetition of major trials is unethical

for saying that one treatment is superior to another. It is usually a matter of judgment whether further trials are needed. The tendency to publish only significant findings and to suppress inconclusive studies certainly distorts in the way that Zelen suggests—a disproportionate number of false-positives to be sorted out by criteria *a* to *g*. There would be less need to exercise this type of judgment if only studies were published that are powerful enough to detect worthwhile and reasonable differences between treatments, irrespective of whether the outcome of such a trial is statistically significant, and if the practice of following the first positive result by a second trial for confirmation was replaced by a scheme that Dr A L Cochrane has

advocated. Instead of one trial followed by another, if the two trials are devised concurrently and reported at about the same time by investigators from different countries then the probability that both will declare false-positives ($p < 0.05$) is about one chance in 400, undistorted by publication or other bias. The method avoids ethical problems, gives a reassuring generality to the conclusions because they have been reached independently by different investigators, and is expedient because it does not delay acceptance of the findings by other doctors. In short, the proposal acknowledges that interpreting statistical significance is not always easy. Planning two trials instead of one is, of course, more expensive, but reluctant approval¹¹ of the results of the Norwegian timolol study and bewilderment over those of the clofibrate trial¹² convince me that vital questions merit this special attention. Perhaps Professor Mitchell's question may be rephrased to ask whether in future major trials should be replicated concurrently so that practical clinical decisions are delayed as little as possible.

References

- Peto R, Doll R. When is significant not significant? *Br Med J* 1977;ii:259.
- Rose G. Beta-blockers in the treatment of myocardial infarction. *Br Med J* 1980;280:1088.
- Altman DG. Statistics and ethics in medical research. Interpreting results. *Br Med J* 1980;281:1612-4.
- Gore SM. Assessing methods—descriptive statistics and graphs. *Br Med J* 1981;283:486-8.
- Gore SM. Mexiletine after myocardial infarction. *Lancet* 1981;i:951.
- MacMahon B, Yen S, Trichopoulos D, Warren K, Nardi G. Coffee and cancer of the pancreas. *N Engl J Med* 1981;304:630-3.
- Peto R. Clinical trial methodology. *Biomedicine Special Issue* 1978;28:24-36.
- The Norwegian Multicenter Study Group. Timolol-induced reduction in mortality and reinfarction in patients surviving acute myocardial infarction. *N Engl J Med* 1981;304:801-7.
- Smith PG, Pike MC, Kinlen LJ, Jones A, Harris R. Contacts between young patients with Hodgkin's disease. *Lancet* 1977;ii:59-62.
- McPherson K. Statistics: the problem of examining accumulating data more than once. *N Engl J Med* 1974;290:501-2.
- Mitchell JRA. Timolol after myocardial infarction: an answer or a new set of questions? *Br Med J* 1981;282:1565-70.
- Committee of Principal Investigators. WHO cooperative trial on primary prevention of ischaemic heart disease using clofibrate to lower serum cholesterol: mortality follow-up. *Lancet* 1980;iii:379-85.

Sheila M Gore, MA, is a statistician in the MRC Biostatistics Unit, Medical Research Council Centre, Hills Road, Cambridge CB2 2QH.

No reprints will be available from the author.

A middle-aged woman with a long history of severe psoriasis recently developed an arthropathy that was treated with benoxaprofen. The arthropathy improved and so did her psoriasis. Is this a recognised effect of the drug?

In many patients psoriasis and psoriatic arthropathy run their somewhat capricious courses independently of each other, although in other patients any treatment that seems primarily to influence one may also help the other. Benoxaprofen is an interesting anti-inflammatory drug¹ as its undoubted anti-inflammatory activity does not seem to correlate with its rather weak antiprostaglandin-synthetase activity. It has been suggested that one facet of its activity may depend on its effect on mononuclear cell migration. There is also an upsurge of interest in monocyte activity in psoriasis,² but so far these converging threads must not lead to premature conclusions. Many drugs have been reported to have dramatic effects in occasional cases of psoriasis and are then found wanting when tried on larger groups of psoriatic patients. Improvement might be due to coincidence, to some direct effect on the main metabolic defect of the psoriasis, or to an effect on side-stream abnormality of importance in that patient at that time. The manufacturers are aware of one or two cases of psoriasis apparently helped by benoxaprofen, but there can be no

justification whatever for its more widespread use until more formal trials have been organised.

- Proceedings of the international symposium on benoxaprofen. *J Rheumatol* 1980;7, suppl 6:1-143.
- Wahba A. Psoriasis: an epidermal disease or a systemic condition. *Int J Dermatol* 1981;20:108-9.

If lettuce and cabbage are good for us, why is not grass a valuable food?

Grass contains much greater quantities of indigestible fibres, such as cellulose, than either cabbage or lettuce and therefore requires much more chewing than most people would have time or inclination for. Improperly chewed vegetable foods can cause intestinal flatulence and colic. The stomach pain caused by eating unripe apples or chunks of raw turnip are familiar examples. Intestinal blockage from large quantities of indigestible fibres is also a possible hazard. Nevertheless, edible protein may be extracted from grass and Pirie¹ has devised such a process. The product has the consistency of a friable cheese and only a slight taste, which it is claimed is not unpleasant. Attempts to introduce such novel protein foods into human diets have met with little success.

- Pirie NW. Leaf protein: a beneficiary of tribulation. *Nature* 1975;253:239-41.