Statistics in Question

SHEILA M GORE

ASSESSING METHODS— CONFIDENCE INTERVALS

Doctors understand so well the importance of checking whether an observed outcome is within the bounds of chance variation that the first two of three pithy definitions of god are as follows: the patient's definition is that the doctor is god, the doctor's definition of god is " $p < 0.05$," while the statistician. . .

Careful reporting does not end at significance testing, especially when the author's summary is that the difference between treatments is not statistically significant ($p > 0.05$).^{1,2} What that amounts to is a statement that the trial results are consistent with there being no difference between treatments, and is not at all the same as saying that there is actually no difference. The distinction becomes clear if the authors report also the range from the smallest to the largest effect of treatment with which the trial data are consistent³—a 95% confidence interval, for example. This range includes zero when the difference between treatments is not statistically significant ($p > 0.05$) but accommodates also real positive and negative differences. If we assumed that the actual treatment effect is outside the quoted 95% limits and were to test the data from that viewpoint we would conclude that the outcome of the trial was improbable. Real differences of such magnitude are therefore confidently, though not definitely, excluded.^{3,4}

This is often a highly informative conclusion but is suppressed by mere testing of significance. Editors should encourage the reporting of confidence intervals for better communication of trial results and so banish for all time the mistake of identifying "not statistically significant" with "no difference."

(10) *Is the treatment of patients with hypertension by trained industrial nurses a practicable alternative to treatment by the patient's family doctor? Four hundred employees with asymptomatic hypertension are randomised to care at the work site or a doctor's care, and the reduction in diastolic pressure is compared.*

The figure gives an idea of the distribution of likely trial outcome if the two methods of treatment are equally successful. What conclusion do you reach if the actual trial outcome is (i) d_1 ; (ii) d_2 ; (iii) d_3 ?

(i) outcome d_1 is consistent with no difference between treatments: 95% confidence interval from -1.1 mm Hg to 2.5 mm Hg straddles zero

(ii) outcome d_2 is improbable ($p < 0.02$) if there is really no difference between treatments: 95% confidence interval from 0.4 mm Hg to 4.0 mm Hg indicates an advantage for treatment at the work site

(iii) outcome d_3 is highly improbable if there is really no difference between treatments: confirmation in the 95% confidence interval, which is from 2.0 mm Hg to 5.6 mm Hg

Horizontal axis is marked off from zero in units of 1 standard error (approx 0.9 mmHg)

- A: Distribution of likely trial outcome when the 2 methods of treating hypertension are equally good
- B: Distribution of likely trial outcome if care at the work site reduces diastolic pressure by a further 3 mmHg
- C: Distribution of likely trial outcome if care at the work site reduces diastolic pressure by a further 2.0 mmHg; outcome d_3 is not improbable $p > 0.05$
- D: Distribution of likely trial outcome if care at the work site reduces diastolic pressure by a further 5.6 mmHg; outcome d_3 is consistent with this distribution

COMMENT

(i) Trial outcome d_1 is consistent with there being a small difference or no difference at all between treatments. From this result we can be 95% confident that reduction in diastolic pressure under a doctor's care is neither better by more than 1.1 mm Hg nor worse by more than 2.5 mm Hg than monitoring by trained industrial nurses. Suppose that industry was prepared to offer treatment at the work site if the true reduction in diastolic pressure was 3 mm Hg more by this method than with a doctor's care. The clinical trial has provided useful information in so far as a differential of 3 mm Hg or more is confidently excluded. Had the industry's threshold been lower—2 mm Hg, say—no practical decision could have been reached, the outcome of the trial being consistent both with no difference and also with a difference of 2 mm Hg.

(ii) Outcome of d_2 is improbable ($p < 0.02$) if there is really no difference between the two methods of treating hypertension. The advantage of care at the work site is confidently estimated as anything from 0.4 mm Hg to 4.0 mm Hg. In particular, the trial outcome is not at odds with a real effect of 3 mm Hg diastolic pressure. If the confidence interval is uncomfortably wide for the decision maker then the trial should have been much larger. Four times as many patients are needed to reduce by half the width of the confidence interval.

(iii) The observed outcome d_3 is highly improbable, given that there is no real difference between care at the work site and a doctor's care. We are 95% confident that the advantage for treatment at the work site lies in a further change in diastolic pressure between 2.0 mm Hg and 5.6 mm Hg. This result would probably lead to providing trained industrial nurses to treat hypertension, but not necessarily so.

The example is based on an actual experiment in occupational health⁵ and was devised so that there was at least a 90% chance of reporting statistical significance ($p < 0.05$) if the real differential was as much as 3 mm Hg. If an industrial concern had been prepared to invest in treating patients with hypertension at the work site they would assuredly have required narrower confidence limits on the potential benefit. Precise estimation of the advantage of one scheme over another can be more persuasive than mere testing of significance. But in certain clinical trials pursuing randomisation beyond the point of having recognised that treatments differ significantly ($p < 0.01$) would be considered unethical.

(11) What do you infer from the approximate 95% confidence intervals shown in the figure below about (i) trial size and (ii) treatments?

(i) the size of trial A is adequate to show convincingly the superiority of chemotherapy, that of trial B is needlessly large

(ii) multiple chemotherapy leads to a highly statistically significant increase in median survival time, which is also clinically relevant

COMMENT

Trial A roughly corresponds to the findings of Mallinson *et al*⁶ who randomised 40 patients with inoperable cancer of the pancreas to multiple chemotherapy versus analgesic treatment (controls). Median survival was 44 weeks in the chemotherapy group, almost five times longer than for patients on the control arm. We can calculate confidence intervals for the ratio of the median survival in the chemotherapy group to the median survival in the control group by first calculating a confidence interval for the logarithm of the ratio—which is approximately normally distributed if the sample size is moderate and survival

Chemotherapy versus analgesic treatment (controls) in patients with inoperable carcinoma of the pancreas: approximate 95% confidence intervals for the ratio of median survival time.

is exponential—and then transforming back. Two scales are shown in the figure. The lower is the exponential of the upper scale: $1 = \exp(0)$ $3.3 = \exp(1.2)$ $11.0 = \exp(2.4)$ and so on. Notice that the confidence interval for $\log_e(\text{ratio})$ is symmetric as usual, from 0.62 above to 0.62 below its midpoint 1.59, but taking exponentials gives an asymmetric confidence interval for the ratio of median survival times, the observed ratio 4.9 being closer to $2.6 = \exp(0.97)$ than it is to $9.2 = \exp(2.21)$. The confidence interval is approximate, given sample size.

Trial A and trial B (hypothetical) both indicate that median survival is significantly longer for patients who received multiple chemotherapy. From trial A we are 95% confident that median survival is at least 2.6 times longer in the chemotherapy group than for control patients who received only analgesics, and might be up to 9.2 times longer. Trial A does not estimate the ratio very precisely but it would have been wrong for the investigators to have continued to admit patients into a trial which had already shown an important effect of treatment—one sufficient to impress not only Dr Mallinson and his colleagues but other doctors as well. Historical data had not given reason to expect a dramatic effect of treatment and initially the authors had thought in terms of a larger trial. Trial B (hypothetical) reports a confidence interval that is only half the width of trial A's but required four times as many deaths to be observed. To have insisted on this greater precision would have been unethical. In the event, a smaller trial was adequate, the plans for a longer series being aborted.

(12) What do you infer from the approximate 95% confidence intervals shown in the figure below about (i) trial size and (ii) treatments?

(i) wide confidence intervals are associated with small trials: to reduce by half or one-third the width of a confidence interval needs four or nine times as many patients respectively;

the size of trial D is much too small; that of trial C is sufficient to give useful information

(ii) from trial C: there is 95% confidence that aspirin does not increase mortality by more than 0.7 deaths in every 100 patients

COMMENT

Trial C in this example—comparison of treatment with aspirin versus placebo in the prevention of secondary myocardial infarction—reproduces the results of Elwood and Sweetnam,⁷ who randomised 1682 patients. The mortality at one year was not significantly lower in the group taking aspirin, zero being

Aspirin versus placebo in the prevention of secondary mortality after myocardial infarction: approximate 95% confidence intervals for difference in mortality rates at one year.

included in the approximate 95% confidence interval for the difference in mortality rates. But the interval conveys more information still. It permits the confident conclusion that an *increase* in mortality attributable to aspirin is no greater than an additional 0.7 deaths in every 100 patients, whereas the benefit could be as much as the prevention of 5.8 deaths. Trial D (hypothetical) can likewise be summarised by saying that treatments do not differ significantly ($p > 0.05$). The summary, however, obscures important differences between trials C and D. Firstly, by comparing the widths of the corresponding confidence

intervals it is clear that the second trial entered only about one-quarter the number of patients randomised by Elwood and Sweetnam: the interval for trial D—from -8.4 to +5.4—having width 13.8 compared with 6.5 for trial C. The second point is that, although both intervals include 0, the confidence limits for trial D succeed only in excluding enormous differences between treatments that would not have been considered realistic anyway. For example, prevention of an excess of 8.4 deaths, more than half the expected mortality, would not have been thought within the capability of any drug a priori. Trial D has not increased our knowledge; trial C was an important and informative contribution, even though it did not show a statistically significant difference between the treatments.

Non-significant differences should not be dismissed just for having failed to reach conventional levels of significance. Confidence limits can advance our understanding; the width of the interval is a guide to how precisely or sensitively a parameter of interest, such as ratio of median survival times or difference in mortality rates, has been estimated. To show convincingly that for practical purposes two treatments are equivalent requires very precise estimation of the differential response, and hence very large numbers of patients.

References

- Rose G. Beta-blockers in immediate treatment of myocardial infarction. *Br Med J* 1980;280:1088.
- Gore SM. Mexiletine after myocardial infarction. *Lancet* 1981;ii:95.
- Gore SM. Assessing methods—art of significance testing. *Br Med J* 1981;283:600-2.
- Armitage P. Statistical inference. *Statistical methods in medical research*. Oxford: Blackwell Scientific Publications, 1977.
- Logan AG, Milne BJ, Achber C, Campbell WP, Haynes RB. Worksite—treatment of hypertension by specially trained nurses: a controlled trial. *Lancet* 1979;ii:1175-8.
- Mallinson CN, Rake MO, Cocking JB, et al. Chemotherapy in pancreatic cancer: results of a controlled, prospective, randomised multicentre trial. *Br Med J* 1980;281:1589-91.
- Elwood PC, Sweetnam PM. Aspirin and secondary mortality after myocardial infarction. *Lancet* 1979;ii:1313-5.

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.

What are the hazards of petrol sniffing?

The toxicity of petrol depends on the origin of the petroleum from which the petrol is derived and the molecular modifications incurred during refining. Saturated hydrocarbons from C4 to C8 have strong narcotic properties and cause nausea, ataxia, and loss of consciousness, whereas unsaturated hydrocarbons in the mixture act as mild anaesthetic agents. The constituents of petrol are highly lipid-soluble and are, therefore, absorbed rapidly from the lungs. Symptoms develop within three to five minutes of inhaling the vapour and 15 to 20 breaths are sufficient to cause intoxication for three to six hours. The euphoria of mild intoxication may be accompanied by nausea and vomiting. After prolonged inhalation, or rapid inhalation of highly concentrated vapour, the sniffer may experience a phase of violent excitement succeeded by loss of consciousness and coma. While unconscious, the subject may exhibit convulsive movements, and his pupils may become fixed and dilated or unequal. Death from inhalation of petrol vapour is rare, but may occur either because of respiratory depression or ventricular fibrillation as a result of cardiac sensitisation to circulating catecholamines by the aromatic fractions of petrol.

The greatest danger from petrol sniffing is related to the possible long-term effects of chronic exposure. Abusers of petrol often inhale the vapour daily for many years. Chronic inhalation may result in loss of appetite and loss of weight, neurasthenia, muscle cramps, and

weakness. Abnormal electroencephalograms have been reported in chronic abusers of petrol, and there is some evidence that permanent neuropsychological damage may develop. Encephalopathy due to organic lead poisoning may also occur in petrol sniffers due to tetraethyl lead added to petrol as an antiknock agent.

Vale JA, Meredith TJ. Poisoning due to aliphatic, aromatic and chlorinated hydrocarbons. In: Vale JA, Meredith TJ, eds. *Poisoning: diagnosis and treatment*. London: Update Books/MTP, 1981.

Is sufficient use made of the antithrombotic properties of aspirin, particularly in the initial treatment of conditions such as acute onset of atrial fibrillation, with its potentially embolic consequences?

The antithrombotic value of aspirin in vivo is the subject of a great deal of debate.¹ Moderate doses (0.5-1.5 g) prevent platelet aggregation in vitro, and, since arterial thrombi are initially composed of platelets, aspirin might be considered a useful agent. What trials there have been suggest that it works better if combined with an anticoagulant than when used alone. In acute atrial fibrillation standard treatment with DC shock and digitalisation, aided if necessary by propranolol, disopyramide, or verapamil, would seem to be preferable, together with attention to the underlying cause of the fibrillation.

¹ Turpie AGG, Hirsh J. Platelet suppressive therapy. *Br Med Bull* 1978;34:183-90.