Statistics in Question

SHEILA M GORE

ASSESSING METHODS— CRITICAL COMMENT

The last article of this series presents mistaken examples from medical journals. Readers are asked to identify oversights in the design of these studies, misuse of statistical methods, or careless interpretation of results. I have not tried to find a comprehensive set of examples—good scientific method, frank reporting, and careful refereeing should make that task difficult—but there are enough cases to remind us to be critical in appraising research papers, even if they have passed an editorial filter. This will ensure that we learn and propagate the best methods. Errors in analysis do not necessarily invalidate conclusions,¹ and careless interpretation can be reviewed by correspondents. Faults in the design of a clinical study are not easily remedied after the event. This is why statisticians primarily emphasise good scientific method.

Having no wish to pillory the papers which inspired these examples, I am grateful to Dr Tony Smith, deputy editor of the *BMJ*, who provided parodies of the original versions. The examples contain the original fallacy, but the medical context has been changed.

(25) *Comment on the analysis (implicit in the figure below) of increased temperature gradient after withdrawal after one year of the β -blocker olalol in 12 out of 29 consecutive patients.*

—17 patients refused to approve withdrawal

—distinct points in the figure do not represent distinct patients

—from the figure it is impossible to deduce for individual patients the change in temperature gradient after withdrawal of olalol

—there is no randomised control group

—temperature gradient was measured knowing that treatment had been withdrawn

—patients who refused were not followed up

—two volunteers were lost to follow-up, with no explanation, at two months

COMMENT

The refusal of 17 patients to take part in the study identifies them as different from the patients who agreed to forego medication. The similarity of the initial mean temperature gradient for volunteers and non-participants does not negate this essential difference. In the figure notice that it would have been better to show the scatter of initial temperature gradient for patients who refused than to present data on the combined group, of which volunteers are a subset.

The distinct points in the figure do not represent distinct patients. Temperature gradients in the right and left leg of the same patient tend to be more similar than such observations on randomly chosen patients. Failure to analyse correctly dependent observations is an error of commission¹: the figure misleads the reader into thinking that the study is more informative than it is. Repeated measurements—as in right and left leg—lead to a better estimation of temperature gradient for a patient, but that is all. They do not increase the number of individuals in the sample.

The change in temperature gradient for individual patients cannot be deduced from the figure because the several measurements made in one patient are not shown as being related. The mean change is, of course, easily computed as the change in means, but the standard error required for Student's paired *t*-test—valid only if the change in temperature gradient is normally distributed—cannot be estimated from the data given. When statistical method and the figure are unrelated, as here, authors should draw attention to this and explain why the data are so represented.

Olalol was withdrawn in all patients who agreed to take part in the study and the authors reported that toe-to-knee temperature gradient increased as a result. The increase cannot logically be attributed to the withdrawal of medication because at the same time as olalol was withdrawn other changes had occurred coincidentally. One particular change was that the patients entered a clinical study, which might in itself account for the observed difference in temperature gradient, making treatment irrelevant. Had patients been randomised to olalol or matched placebo during the study and had there been a significant increase in temperature gradient for the placebo group relative to patients randomised to continued medication, the authors' conclusion could have been confidently made, but not in the absence of a randomised control group.

The patients understood the nature of the study because they gave informed consent, and the assessor was therefore unlikely to have remained in the dark. Knowing that treatment had been withdrawn may have biased the measurement of the temperature gradient. The accusation of bias cannot be refuted. This shows up the experimental design as imprudent.

Not only was there no randomised control group, but the temperature gradient was not measured in the patients who continued to take olalol, having refused its withdrawal. Patients who refuse should be followed up in the same way as patients who give consent. If temperature gradient increased also in patients who persisted with β -blockade, a suspicion that some unexpected time trend might account for the results would be reinforced. A cautionary tale about before-and-after comparisons has been reported by Christie.² Avoid this type of comparison because the inference is uncertain.

Finally, no explanation was given by the authors for the loss to follow-up at two months of two out of the 12 volunteers. Did these two revert to β -blockade and, if so, why?

(26) *Does the advice about drinking fewer cups of tea follow directly from the analysis below?*

—no

—a prospective study in which patients are randomised

Tea drinking and deaths from subarachnoid haemorrhage

Analysis

The overall matched pair risk ratio (R-R) between current tea-drinkers and non-tea-drinkers was 1.9 (95% confidence interval 1.5 to 2.4)

For drinkers of fewer than 2 cups a day the relative risk was 1.2

At a level of 2-4 cups a day the relative risk was 2.3

And for drinkers of 5+ cups a day the relative risk was 4.0

Conclusion

These results suggest that the heaviest drinkers could halve their risk of death from subarachnoid haemorrhage by reducing their tea consumption to an intermediate level; and that benefit of a similar order would be experienced by drinkers of 2-4 cups a day who cut down to less than 2 cups daily

to receive or not receive encouragement to drink fewer cups of tea could give direct evidence

COMMENT

The relative risk of subarachnoid haemorrhage for women who reduce their tea drinking from more than four cups a day to between two and four cups a day may be quite different from the relative risk for women who never drank more than four cups of tea a day. An increased relative risk in association with drinking more cups of tea is not sufficient to determine cause. For instance, women who are at high risk of developing subarachnoid haemorrhage may drink more tea because of this. Reducing their intake would not alter the underlying risk factor. Since the authors do not give the relative risk, even for women who have chosen to drink fewer cups of tea than previously, their conclusion, however plausible, remains mere speculation. The theory that subarachnoid haemorrhage will be less frequent in patients who have been instructed to drink fewer cups of tea can be tested formally by randomising patients to receive or not receive advice about reducing tea consumption.

(27) *Is the evidence in the box below sufficient to recommend the self-treatment of warts with vinegar?*

—no

—readers who tried the method and failed to find relief were not motivated to write about it

COMMENT

The method's success rate could be estimated only if every reader who tried treating warts with vinegar had written to the magazine. Even then the population of patients could not be defined and the rate of spontaneous remission is unknown. This means that the success rate for self-treatment of warts with vinegar cannot be shown to be better than spontaneous remission.

The "experiment" is even more anecdotal because the

Self-treatment of warts with vinegar was described in the medical news column of a women's magazine. Readers who tried the method were encouraged to write and say if it was successful. Thirty letters were received in four months, 25 from readers who said that the method had worked spectacularly. No reader wrote to say that the method was tried but had failed.

magazine will not receive letters from every reader who tried self-treatment with vinegar. I suspect that readers who tried the method and failed to find relief were not motivated to write about it. This being the case, not even the treatment's rate of success could be estimated from the reports that the medical columnist has received.

(28) Does the correspondence in the box below refute the earlier report that survival is significantly longer for women than for men with rectal cancer?

—no

—although the difference in survival is not statistically significant in the second series, median survival times are similar to those in the original report

—statistical significance depends on the number of patients studied as well as on how different the survival of men and women with rectal cancer really is

—a real improvement of 2½ months in median survival may not be enough to justify antiandrogen treatment for men

COMMENT

The authors of the first report on 243 patients were surprised to find that survival was significantly longer for women than for men with rectal cancer and suggested that the observation needed to be checked in other series. If substantiated, a clinical trial of antiandrogen treatment for men patients would follow.

A second series of 72 patients with rectal cancer was reported. The correspondent noted that median survival was not significantly different between the sexes. Median survival times were, however, similar to those in the original report—a point that was not made in the letter to the editor. Statistical sig-

First report: Median survival for 127 women with rectal cancer was significantly longer ($p < 0.01$) than for 116 male patients, half of whom died within 4½ months of diagnosis

Correspondence: The records of 72 unselected patients with rectal cancer were reviewed. There were 39 men and 33 women. Mean age at diagnosis and treatment for metastatic disease were similar; median survival (6 months for women, 4½ months for men) was not significantly different between the sexes

nificance depends on the number of patients studied as well as on how different the survival characteristics of men and women with rectal cancer truly are. If the survival pattern were indeed the same in the two series and survival could be assumed to be exponential, then there is only a 40% chance that by investigating 72 patients the survival difference between men and women would be statistically significant. The second report is therefore not at odds with the outcome in the original series, the second investigation having low power.

Whether an expected improvement of 2½ months in median survival is of sufficient clinical importance to warrant a randomised controlled trial of antiandrogen treatment for men is a decision for doctors to take. Statistical significance is of no account.

(29) Comment on the data presented in the table below and on the comparison of means using Student's paired *t*-test.

—variance is not constant; the three measures of outcome need to be transformed before using classic tests

—for the first two outcomes, response to exercise and the number of anginal attacks in two weeks, the standard deviation increases in proportion to the mean, indicating that a logarithmic transformation is appropriate

—the variance of trinitrin dose increases in proportion to the mean dose and so a square root transformation is needed

—the table does not give the mean and standard deviation of change in outcome, measured within patients on the transformed data, which are needed for paired comparison using Student's *t*-test

COMMENT

Since variance is not constant for any of the three measures of outcome the observations need to be transformed before analysis. A constant ratio of mean to standard deviation indicates

Twenty-eight patients with chronic stable angina entered a randomised double-blind cross-over trial that compared propranolol 80mg daily with placebo. The trial periods were two weeks each. At the end of the cross-over study all patients were put on propranolol 80mg daily and monitored on an open basis for another four weeks, and the outcome at the end of the second fortnight was recorded. The table summarises changes in ST-segment in response to exercise, anginal attacks, and consumption of trinitrin

TABLE Means and standard deviations: 28 patients

Observation	Randomised placebo (2 weeks)		cross-over trial of propranolol (2 weeks)		Open study of propranolol (4 weeks)	
	mean	SD	mean	SD	mean	SD
(Peak ST-segment* / exercise time) x 1000	341	33	199	17	195	17
Anginal attacks per 2 weeks	14.1	5.3	6.7	2.8	4.0	2.2
Trinitrin (mg) per 2 weeks	19.6	5.0	9.7	3.4	4.3	2.1

* lead CM5

a logarithmic transformation. To take response to exercise as an example, the three ratios are similar, being $341/33 = 10.3$; $199/17 = 11.7$; $195/17 = 11.5$. A constant ratio of mean to variance on the other hand shows the need for square root transformation as in the case of trinitrin dose, the ratios of mean to variance being $19.6/25.0 = 0.8$; $9.7/11.6 = 0.8$; $4.3/4.4 = 1.0$. Having obtained as transformed observations for each patient the logarithm of response to exercise, the logarithm of the number of anginal attacks, and the square root of trinitrin dose, the investigator should then check that variance is approximately constant on the transformed scales and that the distribution of within-patient differences (placebo versus propranolol) is approximately Gaussian. Only then should he appeal to Student's *t*-test to assess whether the mean difference deviates significantly from zero. Because it is mean *within-patient difference* which is being tested, the standard deviation which is relevant is the standard deviation of those differences. In addition to the table given the author should report mean within-patient differences on the transformed scales and the corresponding measures of dispersion.

Notice that the within-patient comparison of placebo versus propranolol (four weeks) is a non-randomised comparison and that patients were probably aware of being given propranolol during the open phase of the study. This knowledge may have conditioned their reporting of the frequency of anginal attacks and the dose of trinitrin used.

(30) "*Nepalese hillmen use the right hand to eat food; the food contains possibly protective herbs. Among these people, skin cancers were located significantly more often on the left hand than on the right, and so it follows that the herbs are protective against skin cancers.*" Comment.

—the conclusion does not follow from the evidence

—skin cancers may be more common on the left hand anyway, irrespective of eating habit

—even if left predominance were peculiar to the Nepalese, the herbal hypothesis is not established; a direct test of it could be made, however

COMMENT

The conclusion does not follow from the evidence for two main reasons. Firstly, skin cancers may be more common on the

left hand anyway, so that left predominance is not peculiar to the Nepalese. Secondly, even if left predominance were exaggerated in the Nepalese hillmen relative to other peoples whose diet or eating habit, or both, is different, the herbal hypothesis would not be firmly established. Other explanations come readily to mind. Beware of speculative inference. A direct test—of whether the handling of food that contains certain herbs protects against skin cancer—might be impracticable. It would at least entail long-term follow-up of subjects who had been randomised to particular combinations of diet (including or excluding certain herbs) and eating habit (use left hand, right hand, knife and fork). Comparing the incidence and time to appearance of skin cancers would be appropriate.

This is the final article in the series Statistics in Question.

I gladly record my thanks to Dr Ian Sutherland, director, MRC Biostatistics Unit, and to Professor Calbert I Phillips, department of ophthalmology, University of Edinburgh, for constructive criticism of all of these articles.

I am grateful to Dr Stuart J Pocock, Royal Free Hospital, London, Dr Tony L Johnson, Ms Helen Tate, and Mr Derek Lowe, MRC Biostatistics Unit, for valuable commentary on particular articles, and to the many research workers who generously allowed me to reproduce figures.

I am indebted to Sue Burkhart, staff editor on the *BMJ*, who liaised with me in co-ordinating the series, and outlawed as many unhelpful statistical terms as was possible. The challenge of devising a suitable layout for a question and answer series was willingly and, I think, admirably met by the editorial staff of the *BMJ* in collaboration with Mr Derek Virtue, their medical artist, whom I thank.

Miss Sally Stephenson and Mrs Beryl Smith, MRC Biostatistics Unit, ensured the smooth preparation of these articles by the efficiency and patience of their secretarial help.

References

- Gore SM, Jones IG, Rytter EC. Misuse of statistical methods: critical assessment of articles in *BMJ* from January to March 1976. *Br Med J* 1977;i:85-7.
- Christie D. Before-and-after comparisons: a cautionary tale. *Br Med J* 1979;iii:85-7.

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.

In view of the increased safety of the enteric-coated preparations of non-steroidal anti-inflammatory drugs, and also aspirin, should we be prescribing so many of these drugs in the usual form which allows them access to the gastric mucosa with the added risk which that entails?

The mechanism by which the non-steroidal anti-inflammatory drugs produce gastric mucosal ulceration is not entirely clear but it is thought to include inhibition of gastric cyclo-oxygenase. This occurs to some extent whatever the route of administration¹ and will therefore not be eliminated by enteric coating, though it may be reduced. Apart from aspirin (Nu-seals, Caprin, etc), the only non-steroidal anti-inflammatory drugs available as enteric-coated preparations are phenylbutazone (Butacote), oxyphenbutazone (Tandacote), and diclofenac (Voltarol). Enteric-coated preparations of aspirin improve gastric tolerability and reduce the incidence of mucosal ulceration and bleeding.^{2,3} Butacote and Tandacote are also better tolerated than the uncoated preparations but whether or not they reduce the risk of ulceration is not known. Gastric intolerance, ulceration, and bleeding are less of a problem with the newer anti-inflammatory drugs and Voltarol does not seem to be any better in this respect than other drugs, such as ibuprofen, naproxen, etc. Doubts have been expressed about the bioavailability of enteric-coated preparations

but new preparations of enteric-coated aspirin, including Nu-seals, seem to be reliably absorbed, as do Butacote and Voltarol. Enteric-coated preparations are, however, more slowly absorbed and are therefore unsuitable for single-dose analgesic use. Anti-inflammatory doses of aspirin (over 3 g daily) are associated with a high incidence of gastrointestinal side effects, and for patients unable to tolerate such doses enteric-coated preparations may be useful if the drug is otherwise effective. They may possibly also be preferable if high doses of aspirin are required for a patient with a history of peptic ulceration. They are more expensive than soluble aspirin and should not be used routinely. The incidence of gastrointestinal side effects is much lower with the newer anti-inflammatory drugs, and I doubt whether enteric-coated preparations would do much to improve their tolerability and safety.—LINDA BEELEY, consultant clinical pharmacologist, Birmingham.

¹ Duggan JM. The pathogenesis of the aspirin-related gastric lesion. *J R Coll Physicians Lond* 1981;15:117-8.

² Lanza FL, Royer GL, Nelson RS. Endoscopic evaluation of the effects of aspirin, buffered aspirin and enteric coated aspirin on gastric and duodenal mucosa. *N Engl J Med* 1980;303:136-8.

³ Hofstetzer JW, Silvano GR, Burks M, Ivey K. Comparison of the effects of regular and enteric-coated aspirin on gastroduodenal mucosa of man. *Lancet* 1980;ii:609-12.