of the urinary tract, and often a micturating cystogram, though not invariably during the first infection. We appreciate that if the necessary expertise is not available to obtain a high standard of ultrasound then an intravenous urogram is still necessary in this group. If initial studies show no abnormality then the child is simply followed up. If they show abnormality radionuclide tests and other examinations are indicated as for the previous group.

Over 5 years (fig 3)—In this group we believe that progressive damage from reflux nephropathy is most unlikely, so we do not feel the need to diagnose reflux during the first screening procedures. At present in this group we perform an ultrasound and plain *x* ray examination of the abdomen only. If appearances are entirely normal we follow up the child, and perhaps go on to further investigations at the time of the next infection. If the initial tests show an abnormality the child would undergo either an intravenous urogram or micturating cystourethrogram or both as indicated.

Urinary infection is common in children. Correctable structural causes are few but important. We think that we should make use of new imaging techniques if they are kind and effective, but only to look for what is correctable and important.

References

1 Bailey RR. An overview of reflux nephropathy. In: Hodson J, Kincaid-Smith P, eds. *Reflux nephropathy*. New York: Masson Publishing USA Inc, 1979:3-13.

2 Birmingham Reflux Study Group. Prospective trial of operative versus non-operative treatment of severe vesicoureteric reflux: two years' observation in 96 children. *Br Med J* 1983;**287**:171-4.

3 Kincaid-Smith P. Reflux nephropathy. *Br Med J* 1983;**286**:2002-3.

4 Lanning P, Seppänen U, Huttunen N-P, Uhari M. Prediction of vesicoureteral reflux in children from intravenous urography films. *Clin Radiol* 1979;**30**:67-70.

5 Cavanagh PM, Sherwood T. Too many cystograms in the investigation of urinary tract infection in children? *Br J Urol* 1983;**55**:217-9.

(*Accepted 17 November 1983*)

Mathematics in Medicine

Statistical ritual in clinical journals: is there a cure?—I

DONALD MAINLAND

A disturbing verdict

"The presentation of variability in medical journals is a shambles." This verdict from a Medical Research Council statistician[1] appeared 55 years after the first edition of R A Fisher's *Statistical Methods for Research Workers*, the main source of the statistical arithmetic now widespread in medical journals: standard deviations, standard errors of the mean, *r*'s, significance tests (especially χ^2 and *t*) with the resulting P's and *p*'s,* and, more recently, confidence intervals. The verdict seems astonishing because the main reason for using statistical methods is variability; and it is especially disturbing, even conscience pricking, to one who soon after *Statistical Methods* appeared started applying it to his own research,[2] mostly using "significance" tests which, to avoid serious ambiguity, should have been called "chance frequency" tests. During 30 years or so after this I tried to propagate the elementary ideas of Fisher's *Design of Experiments*, which in 1935 presented randomisation as the logical basis for significance tests, to replace the mythical doctrine that a study group and its control should be "alike in all respects" except the factor or factors under investigation.

*The author is following Fisher's symbolism where P is used to denote the probability (random frequency) of an observation plus the probability of rarer observations in the same or opposite tail, and *p* represents the probability of a single class of observations.

Kent Hills, Kent, Connecticut 06757, USA

DONALD MAINLAND, MB, DSC, retired professor of anatomy and medical statistics

Correspondence to: Kent Hills, Apt 3B1, Kent, CT 06757, USA.

Thinking versus arithmetic

One of the earliest clinical applications of Fisher's *Design of Experiments* was the trial of streptomycin in pulmonary tuberculosis, which emphasised not arithmetical but statistical *thinking*—concern with variability and with risks of bias throughout the planning and performance of a study.[3] After the 'forties controlled trials (of diverse quality) spread widely, but statistical arithmetic spread more abundantly and less rationally; and surveys of medical journals have shown that "in about half of all published articles there are statistical errors," including misuse of arithmetical techniques.[4-9] Fisher himself came to deplore how often his own methods were applied thoughtlessly as cookbook solutions.[10] Among reasons suggested for this state of affairs are: emphasis on techniques in elementary textbooks[11]; lack of contact with medical problems during statisticians' training[8]; too much "mathematistry" (development of theory for theory's sake) during the education of consultant statisticians for all aspects of science.[12 13]

In attempts to improve the position the *BMJ* has issued four series of articles, later published in book form,[14-16] and a set of guidelines which should help in planning and reporting research[17]; but I wonder whether any of these publications will have much effect on the arithmetical ritual.

A protest

Referring to the *BMJ*'s instructional articles a letter from a department of pathology asked why one should "try to fit recalcitrant numbers into some analysis which will confirm what one knows in one's bones to be true."[18] The letter quoted two papers in which non-significant differences were apparently counted as "real," and asked: "If we are to be allowed to ignore

statistics after we have gone through the laborious task of calculating them, could we not skip them altogether?"

The two examples will be discussed later, but in her own work the critic would obviously find congenial the test christened by Berkson, the Mayo Clinic statistician, "the interocular traumatic test"—the verdict hits you between the eyes. In such cases the important question is: What, apart from wishful thinking, is the source of the "intraosseous" or "interocular" conviction of the truth? The answer might suggest how to escape statistical ritual in less obvious cases.

Seeking escape routes

Looking for ways to avoid ritual, either unnecessary or inappropriate, I examined many articles in the 1979-82 issues of the *BMJ*, searching for the authors' interpretations of their arithmetic, and then asking: Could they have profitably skipped some of the ritual? A few of the resulting soliloquies will be mentioned below, in the perhaps naive hope that a more penetrating study will be performed on their own data by medical research workers who are acquainted with the implications and limitations of the arithmetic—perhaps helped by a very open minded statistician.

Absence of raw data

My explorations were somewhat hindered because most authors had disregarded the old adage that one table of raw data is worth more than a half dozen tables of derived values. I thought of the loss to other readers who wished to form their own opinions from the recorded observations, perhaps to answer questions not raised by the authors, and to seek exceptions and individual peculiarities, so fundamental in medicine, in contrast to the statisticians' traditional concern with groups. Even with several hundred subjects the individual data can be shown photographically or in miniprint for readers to scrutinise. But how many would do that? Nowadays, even investigators do not necessarily do it. With computers they can produce reports without really looking at their raw data.[19]

Randomisation and testing

Each patient in a trial is a kind of algebraic sum of factors (biases), each pushing him either towards "success" or "failure" —factors present at the start and events, in addition to the treatments under test, that will occur during the trial. The randomisation assigns these net biases to one or other treatment group, and at the end the inference is a dichotomy: either the treatments or the randomisation caused the intergroup difference in outcome—provided that nothing, known or unknown, except the treatments, interfered with the effect of the randomisation.

The most obvious way of finding how far the randomisation might be responsible is by a randomisation trial—random assignment, say 1000 times, of each patient's outcome measurements to arbitrarily labelled classes A or B. With computers this can now be done, but is still too expensive for most studies; so we still use the mathematically invented significance tests and face, or more often ignore, the problems that will be discussed in part II.

Many investigations in medicine are observational studies (surveys) because it is impossible or unethical to assign experimentally, and randomly, the factors under investigation. The inference after a test is then not a simple dichotomy. The real cause of a significant difference may be a factor or factors undetectable in the data. Are the tests, therefore, as valuable as many addicts seem to think?

Possible test reduction

Surprisingly, we may obtain help in test reduction from Fisher, whose writings show that he did not, like addicts today, consider tests as mathematical yes/no proofs.[20] In section 7 of *The Design of Experiments* he wrote: "We may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result."[21] If the reference to statistical significance is replaced by "rarely fail to show a result in the same direction," does not the statement represent the traditional method of establishing scientific conclusions? Why, then, should significance tests be considered essential after every individual experiment or survey? If we always remembered that individual studies are usually just contributions to a pool of information, significance tests ought to lose much of their apparent importance, and perhaps would often disappear. "Pooling" does not imply elaborate and probably controversial efforts to combine studies numerically; it implies thoughtful weighing of evidence.

Playing safe

A journal referee recently condemned failure to apply statistical tests when they "ought" to be applied. Because of the uncertainties inherent in test verdicts, it is not so easy to define the circumstances that justify the imperative. It would be undiplomatic for investigators to turn violently against the common custom, but even if they produced P's merely as passports to publication they could try to counteract the abominable ambiguity of "significant." They should ask: Was this test superfluous? what did the result of that test really tell me? The answers would be educational to readers.

For example, after a plasma exchange using a fluid of lower viscosity than normal, the plasma viscosity of the recipients was diminished ($P < 0{\cdot}001$). Was the test necessary, or was the amount of the reduction the important figure?

Six months after the performance of a widely used operation for the relief of cardiac pain not responsive to medical treatment most of the patients could perform work that was previously impossible. A significance test showed that this would very rarely happen by chance.

Muddled interpretations

NOT SIGNIFICANT

Even after half a century of warning, some investigators apparently still equate not significant to not real. Others equate it to insignificant (trivial, of no practical importance). Perhaps they have seen experts, without mentioning the warnings, proclaim that a study group and control group are comparable because there are no statistically significant differences in various observed features (age, sex, etc).

Ambiguity of interpretation can explain the confusion in the pathologist's letter regarding two non-significant differences that the authors apparently considered real—an interhospital difference in neonatal mortality and a decrease in blood pressure under treatment.[18] The reports were not explicit, but the authors of one of them stated that they had not searched for causes of the differences in mortality, and the other author clearly expected a known antihypertensive drug to produce a fall in blood pressure but found that the decrease was not as great as with a combination of that drug and another weaker one.[22 23]

Test verdicts should not supersede background knowledge. In eight patients with diabetes who showed evidence of β pancreatic cell activity the mean duration of the disease was 3·2 years; in seven others who showed no such evidence the

mean duration was 6·3 years. The difference in duration was marked "NS," but it would be more in line with current knowledge to accept it as probably real.

P, p, AND NS

Significance tests entail "probability" and we should all know what the word and its symbols mean in that context. Therefore I was sorry to see that the *BMJ* guidelines,[17] following the Vancouver rules, used p (roman lower case) instead of P (roman capital) which for years, in accordance with Fisher's symbolism, has indicated the probability (random frequency) of an observation plus the probability of rarer observations in the same or opposite tail, whereas p (lower case italics) represents the probability of a single class of observations.

When a reader sees the abbreviation NS he can usually assume that the cut off point was 0·05, but he does not know whether P was 0·70 or 0·07, and for one who is accumulating independent information on the same topic the individual P's can be helpful.

At the other extreme, P *greater* than 0·95 looks like strong evidence in favour of the null (no difference) hypothesis, but this is fallacious.[24] After a comparison of treatment three P values from χ^2, 0·96, 0·97, and 0·98, were marked NS. The last figure, for example, would imply that if the randomisation were the only cause of the difference in outcome not more than 2% of randomisations would cause as great a *similarity* as the one observed. The figures were not due to an inaccurate formula, and the trial was carefully conducted but could not be blindfold; so possibly the observers or patients, or both, tended to minimise the differences in outcome.

More fundamental analysis

Even if we apply statistical tests, more important analysis is more old fashioned. It includes detailed scrutiny of what went on in the study and what has, or may have, gone wrong. It entails also thorough consideration of the agreement and especially the disagreement of our observations with previous reports and relevant theories.

References

1 Altman DG. Statistics and ethics in medical research. *Br Med J* 1980; **281**:1542-4.
2 Mainland D. Personal view. *Br Med J* 1980;i:1269.
3 Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *Br Med J* 1948;ii:769-82.
4 Schor S, Karten I. Statistical evaluation of medical journal manuscripts. *JAMA* 1966;**195**:1123-8.
5 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.
6 Glantz SA. Biostatistics: how to detect, correct and prevent errors in the medical literature. *Circulation* 1980;**61**:1-7.
7 Wallenstein S, Zucker CL, Fleiss JL. Some statistical methods useful in circulation research. *Circ Res* 1980;**47**:1-9.
8 Altman DG. Statistics in medical journals. *Statistics in Medicine* 1982;**1**: 59-71.
9 Gardner MJ, Altman DG, Jones DR, Machin D. Is the statistical assessment of papers submitted to the "British Medical Journal" effective? *Br Med J* 1983;**286**:1485-8.
10 Box JF. *R. A. Fisher: the life of a scientist*. New York: Wiley, 1978.
11 Mainland D. Medical statistics—suggestions for the evaluation of introductory textbooks. *J Chronic Dis* 1983;**36**:345-51.
12 Box GEP. Science and statistics. *Journal of the American Statistical Association* 1976;**71**:791-9.
13 Feinstein AR. Clinical biostatistics XL: stochastic significance, consistency, apposite data, and some other remedies for the intellectual pollutants of statistical vocabulary. *Clin Pharmacol Ther* 1977;**22**: 113-23.
14 Swinscow TDV. *Statistics at square one*. London: British Medical Association, 1976.
15 Rose G, Barker DJP. *Epidemiology for the uninitiated*. London: British Medical Association, 1979.
16 Gore SM, Altman DG. *Statistics in practice*. London: British Medical Association, 1982.
17 Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. *Br Med J* 1983;**286**:1489-93.
18 Lester E. Statistics in medical research. *Br Med J* 1980;**281**:1568.
19 Altman DG. Statistics and ethics in medical research. *Br Med J* 1980; **281**:1399-1401.
20 Mainland D. Medical statistics—thinking vs arithmetic. *J Chronic Dis* 1982;**35**:413-7.
21 Fisher RA. *The design of experiments*. 3rd ed. Edinburgh and London: Oliver and Boyd, 1942.
22 Steiner ES, Sanders EM, Phillips ECK, Maddock CR. Very low birth weight children at school age: comparison of neonatal management methods. *Br Med J* 1980;**281**:1237-40.
23 Seedat YK. Trial of atenolol and chlorthalidone for hypertension in black South Africans. *Br Med J* 1980;**281**:1241-3.
24 Fisher RA. *Statistical methods for research workers*. 3rd ed. Edinburgh and London: Oliver and Boyd, 1930.

(*Accepted 29 November 1983*)

Indian reflections: doctor's dilemma

Complaints against doctors and hospital services are quite frequent now. Death or suffering has been attributed to the callousness or negligence of a doctor or to the inefficiency of hospital services. Inquiries have been started to apportion blame and a doctor finds himself in the dock to answer charges. He is today a maligned man.

What has he to say about it? The qualities which are expected in him are well known. Though he is also susceptible to all the evils which can befall flesh and blood his image has been placed on a high pedestal. Time and again he is reminded that he belongs to a "noble profession" so he has to be different from others. In spite of all this a doctor has no illusions about himself or his profession. He claims no "aura" or "halo." Like other professional people he considers himself a technical person. But with a difference, because he cannot really claim to possess all the knowledge of the complex human organism. Man has complete mastery of the wonderful spacecraft, satellite and rocket, which are his creation but how can he say the same thing about a living being, the components of which he has neither created nor even assembled? As a matter of fact the higher a doctor rises in the rungs of the profession the more bewildered he is to discover the areas of darkness in his knowledge.

Like mathematics medicine is not an exact or accurate science. No two identical cases are exactly alike even if they are suffering from the same illness. It is in medical practice one finds two plus two may not always be four. Unexpected reactions and unpredictable results have happened in the hands of the most competent. Such incidents are dubbed as miracle or disaster by the laity as the case may be. As such the proverb "one man's meat is another man's poison," in a broad sense has to be borne in mind by the doctor.

It has been said that the practice of medicine is an art, but to some it is a craft. Probably it is a mixture of science, art, and craft appropriately blended and dispensed taking into consideration the "psyche" and "soma-atos" (Greek, body) of the patient who seeks advice or treatment. A self employed person will not usually like to spend more than is necessary to cure his illness, whereas a person in service, if he is entitled to medical reimbursement benefits, has a different view. Even for a minor ailment he may not appreciate any economy on the part of the doctor in his treatment. On the plea of a rapid recovery he may suggest or even demand what he strictly does not require. An educated person may show off his medical knowledge and suggest his own treatment to the doctor. A doctor who lets his own superior judgment prevail will fail to impress such a patient. The dissatisfied patient may then seek the treatment of his choice elsewhere. He may even fall into the hands of a quack. Is it any wonder then that potentially dangerous drugs which have flooded the market are being used indiscriminately?

Scientific medical practice cannot continue if a patient does not believe in it or has no patience for it. The alternative then left to a doctor is the practice of art or craft in order to maintain his popularity and clientele.—S K MAJUMDAR, director, HHRD Medical Trust, Jodhpur, Rajasthan, India.