Common sense, nonsense and statistics

Nick Lane PhD

J R Soc Med 1999;**92**:202–205

The media has recently been echoing a claim that medical statistics too frequently inspire cries of 'breakthrough!' later exposed in the cold light of experience, as nothing but mirages. The arbitrary parameters and potential subjectivity of conventional medical statistics have been debated, without consensus, in most of the major journals. The question stands: do we allow conventional medical statistics to bias our interpretation of the true significance of medical research?

Undoubtedly, some form of statistics is needed to help interpret the results of clinical trials. The problem is the cut-off point: what qualifies as an important finding? For years, conference halls have rung with sceptical voices questioning the clinical relevance of statistically significant data. Neither *P* values nor confidence intervals seem consistently reliable as a guide to clinical significance. Moreover, the assumptions underlying parametric significance tests (such as equality of variance in group comparisons) may be untrue of some studies. A systematic approach often put forward to correct for subjectivity bias (prior probability) among clinical investigators is Bayesian analysis. We are told that elimination of subjectivity by use of Bayesian inference paves the way to truly objective, evidence-based medical practice. Yet who but a statistically minded minority can begin to interpret Bayesian analysis?

Reading various exchanges of letters in the medical journals on statistical models, I am struck by the fact that almost all are written by biostatisticians. The opinions of practising physicians are deafeningly silent. In our mounting enthusiasm for evidence-based medicine, we may be giving too much weight to clumsy statistical formulations that should not hinge our judgment of clinical relevance. In this article I suggest, firstly, that conventional medical statistics are better than more sophisticated alternatives since they are better understood by practising physicians; and, secondly, that the problems we have with statistics derive less from our methodology than from our stiff conventions of interpretation.

'GUSTO' REVISITED

A once-controversial example illustrates my first point—GUSTO, the largest trial of thrombolysis in acute myocardial infarction. No seminal trial ever generated more animosity or was a fitter target for the Reverend Bayes. The GUSTO investigators sought to prove that modern, high-tech, alteplase is better than dreary old streptokinase. True to its colours, GUSTO found a 'significant' 1% difference in mortality in favour of alteplase[1], immediately provoking a storm of criticism. The criticisms are familiar to all those who have followed the fortunes of thrombolysis. GUSTO was an open-label trial, sponsored by the manufacturers of alteplase, in which a conventional regimen of streptokinase was compared with an optimized regimen of alteplase. More patients receiving alteplase had 'rescue' angioplasty, artificially lowering mortality in this group. Many American investigators were unfamiliar with streptokinase, and terminated infusions in patients who developed hypotension; European investigators, who were generally more familiar with streptokinase, took appropriate precautions and continued infusions. The GUSTO investigators themselves fuelled protests by publishing two separate amendments to their data. In short, few trials have embraced prior belief so publicly. And inevitably, when subjected to Bayesian analysis, all differences between streptokinase and alteplase evaporated—except for an excess of haemorrhage in patients treated with alteplase[2].

The storm died away. The issues gravitated back, as they usually do, to price and prejudice. The great majority of physicians who can afford it now use alteplase; a stubborn minority persist in their view that streptokinase is just as effective and ten times cheaper. Sadly, many eligible patients do not receive any thrombolytic at all. The balanced Bayesian interpretation is ignored. Why? I would suggest that the most appropriate conclusion to draw from GUSTO and its aftermath is that most physicians just prefer alteplase. The statistical difference between the two thrombolytics is perceived as inconsequential. Bayesian analysis adds little, and indeed the zeal of its exponents raises the question: are the advocates of Bayesian inference themselves biased? Should we perhaps apply Bayesian analysis to the Bayesians?

Alteplase is preferred not because it is better but because it is less worrying. Patients receiving alteplase tend to revive quickly and 'look' better. A few will die of cerebral haemorrhage. Only rarely is there an extended period of anxiety during which physicians, nurses and

University Department of Surgery, Royal Free and University College Medical School, Pond Street, London NW3 2QG, UK

relatives must keep vigil over the patient as is so often the case with streptokinase infusion. The point I want to make is that physicians' belief in, or comfort with, a therapy is a powerful force that should not be dissipated by a misplaced striving for 'objectivity'. If there is a clear difference between two treatments, the medical bandwaggon will always trundle onwards; but if two treatments are so similar that Bayesian analysis is required to distinguish between them, then most physicians will prefer to make up their own minds. Standard statistical methods offer a simple and intelligible yardstick against which physicians can form their own judgment of best clinical practice without the interpretations of middleman. GK Chesterton once complained that literary critics had robbed modern poets of their popular voice[3]: by leaping to praise or condemn obscure lines, these interpreters saved our budding poets the trouble of ever being understood. Today we face a similar danger in medicine—that of excluding clear meaning with sophisticated statistical clutter.

POMP AND CIRCUMSTANCE

Yet the system we have inherited is plainly flawed. We pay lip service to clinical meaning over statistical constructions, but are nevertheless guilty of worshipping the *P* god. That magical phrase 'statistically significant' is still the password to publication. Clinical research is expected to be targeted to practical goals. Academic funding, whether from grants bodies, institutions or pharmaceutical companies, is conditional on a distinguished record of publications, which in turn depends on researchers churning out statistically significant data. No wonder Bayesian analysis has not won widespread popularity—it undermines the very livelihood of medical researchers. While few people fool themselves over the value of their own results, researchers are under relentless pressure to sell their data to journals. The sales spiel we use is a dazzling salvo of technical jargon, third-person distance, relative risk ratios and statistical wizardry—scientific pomp and circumstance. Times may be changing, but primary research papers are still cloaked in a pretence of objectivity that contrasts with the frankness of conference platforms, or the letters pages, editorials and commentaries of the journals. To distort Henry Ford, the verbal agreements of scientists are worth more than the papers they are written up in.

What is at stake here is that old bugbear, the 'scientific method'. Medawar[4] argued that the very structure of scientific papers betrays a vain fallacy—that science proceeds by induction rather than deduction. The process of induction requires a prospective hypothesis, which is tested systematically and accepted or rejected on grounds of statistical probability. What really happens? According to Medawar science proceeds in a much more haphazard fashion. We struggle to understand by fabricating a story. We dream up experiments that might help us fill in the gaps in our story, only to discover that the story is not as simple as we imagined. In the light of our findings, we piece together a new story that seems to fit, fleshing it out with a retrospective hypothesis. Finally, we pretend that the hypothesis was actually prospective and write it up for publication, replete with the 'objective' trappings of the scientific method.

I think that most laboratory-based scientists would recognize in this cartoon a kernel of their daily experience, although I admit my description is a far cry from the methodology of the randomized clinical trial. I do not want to suggest that we should reconsider the design of clinical trials: on the contrary, a prospective hypothesis is necessary to estimate sample size, study duration or probable outcome; and clinical trials are cumbersome beasts that cannot mimic the quicksilver twists of laboratory research. But I do think that, having once strapped on the statistical straitjacket, we have grown unwilling to admit that freedom of movement is even possible. Having formulated a prospective hypothesis, we restrict our analysis to a narrow interpretation of predefined endpoints. We are easily embarrassed by arguments based on common sense and overzealous to shore up claims with statistical 'proof'. Our enthusiasm for the scientific method may lead us to overlook or even dismiss robust clinical trends that need no more defence than common sense.

INTERFERING WITH INTERFERON

Consider the case of interferon beta in the treatment of multiple sclerosis. Interferon beta 1b (IFNβ1b) was the first drug to have a real effect on the progression of relapsing–remitting multiple sclerosis. In a pivotal trial completed in 1993, IFNβ1b was shown to reduce lesion burden and almost eliminate inflammatory activity in new lesions[5]. These surrogate markers corresponded well to a reduction in the number and severity of clinical exacerbations, and correlated with a slower progression of disability; but the effect on disability did not reach statistical significance. Since most patients recruited to the trial were already mildly disabled, and since disability tends to progress slowly at this stage of disease, common sense alone suggests that a longer trial would have detected a more substantial effect on disability. But common sense is sacrificed on the altar of statistics: the statistical shortcomings of the trial have been used instead to argue that IFNβ1b is of doubtful efficacy. Enter interferon beta 1a (IFNβ1a). With the benefit of hindsight, patients with minimal disability were recruited in a pivotal trial completed in 1996. This patient cohort progressed rapidly through the early stages of disease, and IFNβa duly demonstrated a 'significant' effect on the progression of disability. As a result, IFNβ1a quickly captured 80% of the US market, despite equivocal effects on lesion burden and activity.

Now I can find no data to suggest a discrepancy in the biological activities of IFNβ1b and IFNβ1a; the greatest difference lies in their dosing regimens. IFNβ1b is given subcutaneously every other day, IFNβ1a by intramuscular injection once a week. We might be forgiven for assuming that these regimens are based on meticulous dose-ranging studies, but we would be wrong: the meagre dose-ranging studies published for IFNβ1a are rather more supportive of a twice-weekly schedule[7]. This higher dose was in fact neglected for fear of unacceptable side effects. When considered in this context, though, the equivocal effects of IFNβ1b on lesion burden and activity are no longer surprising, while the 'significant' effect on disability can be safely ascribed to careful patient selection. I am forced to conclude not that one interferon is better than another, but that—given their disparity in market share—an over-rigorous application of statistics has served to mislead both doctors and patients.

TRIAL AND ERROR

Another example, still rumbling with controversy, is the use of thrombolytics in acute ischaemic stroke. Alteplase was first tested in stroke in ECASS I[8]. The results suggested that the drug was effective in carefully selected patients, but caused potentially fatal bleeding in patients at high risk of haemorrhagic transformation. Alteplase was later approved for treating stroke in the US on the basis of the exemplary NINDS trial[9]. Conducted in eight specialist centres, NINDS showed that patients treated with alteplase within three hours of the onset of acute ischaemic stroke were at least 15% more likely to make a good recovery than patients receiving placebo. Although the risk of cerebral haemorrhage was tenfold higher with alteplase, mortality was similar in the two groups. The stumbling blocks were rather the short time-window for treatment and the narrow margin for error: in recruiting 624 patients over three years, more than seventeen thousand patients had to be screened. The implication that only a few people will benefit from thrombolysis is borne out by the limited sales of alteplase in the US since it was licensed.

High hopes rested on ECASS II[10]. The trial design assimilated lessons learnt from ECASS I and NINDS, setting the time-window at a more realistic six hours and addressing the complexities of interpreting early computed tomographic scans. On the face of it, the results were disappointing. The prespecified primary endpoint—modified Rankin score dichotomized for favourable (0–1) or unfavourable (2–6) outcomes—did not achieve statistical significance, despite a 3.7% absolute difference in favour of alteplase. Secondary endpoints also consistently favoured alteplase without reaching statistical significance. The problem was at least partly selection bias. Most patients had suffered minor strokes and stood a good chance of full recovery irrespective of their treatment. In support of this interpretation, mortality was very low in both groups (10.6%).

According to statistical convention, ECAS II was unarguably negative. An accompanying editorial in *The Lancet* proclaimed that alteplase was not yet proved, and called for further trials to define which patients would gain most benefit from alteplase[11]. But should we really interpret ECAS II as negative? A *post-hoc* analysis of Rankin scores dichotomized for independence (0–2) or dependency and death (3–6) showed a significant absolute difference of 8.4% in favour of alteplase[10]. A charge of *post-hocery* is misleading. The investigators simply over-estimated the efficacy of alteplase, and therefore the power calculations for ECASS II, on the basis of the strong NINDS data. The absence of a 10% efficacy difference does not mean that smaller differences are clinically meaningless, just that a new trial would have to be larger, slower and more expensive to detect them. Is this necessary? The investigators themselves suggest their data should be interpreted in the light of earlier experience, and conclude that alteplase should be approved for routine use in specialist centres. If I had an ischaemic stroke, I would certainly want to be treated with alteplase by specialists trained in stroke management.

COURTING THE TRUTH

Modern 'evidence-based' medicine has some unsettling parallels with the English legal system. New drugs are considered innocent of efficacy until proved guilty of an effect. As in the courts, our scientific jurors may consider only the evidence placed before them. Character witnesses or circumstances carry little weight compared with an accumulation of 'hard' statistical evidence. Inevitably there are miscarriages of justice. In scientific medicine, this means that some drugs backed by an accumulation of persuasive but circumstantial evidence do not enter clinical practice until technically 'proved'. In the case of IFNβ1b, a trial of the drug in patients with secondary progressive multiple sclerosis (which is, if anything, harder to treat than the relapsing–remitting form) was recently halted early after interim results showed that IFNβ1b indisputably slowed the progression of disability[12]. It is hard to be surprised, and one wonders how many patients have been denied the hope of therapy on what amounts to a technicality. The examples of interferon beta and alteplase turn the media condemnation of medical statistics on its head. Rather than inspiring cries of 'breakthrough' in the medical press, our inflexible use of statistics conceals a deeper and more troubling undercurrent—the rationing of healthcare resources by application of misleading statistical arguments.

REFERENCES

1 The GUSTO Investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. *N Eng J Med* 1993;**329**:673–82

2 Brophy JM, Joseph L. Placing trials in context using Bayesian analysis. GUSTO revisited by the Reverend Bayes. *JAMA* 1995;**273**:871–5

3 Chesterton GK. *The Middleman in Poetry. Essays and Poems.* London: Penguin, 1958

4 Medawar P. Is the scientific paper a fraud? *The Strange Case of the Spotted Mice and Other Classic Essays on Science.* Oxford: Oxford University Press, 1996

5 IFNB Multiple Sclerosis Study Group. Interferon beta-1b is effective in relapsing–remitting multiple sclerosis. *Neurology* 1993;**43**:655–61

6 Jacobs LD, Cookfair DL, Rudick RA, *et al.* Intramuscular interferon beta-1a for disease progression relapsing multiple sclerosis. *Ann Neurol* 1996;**39**:285–94

7 Jacobs LD, Munschauer FE. Treatment of multiple sclerosis with interferons. In: Rudick RA, Goodkin DE, eds. *Treatment of Multiple Sclerosis: Trial Design, Results and Future Perspectives.* London: Springer, 1992:223–50

8 Hacke W, Kaste M, Fieschi C, *et al.* Intravenous thrombolysis with recombinant tissue plasminogen activator for acute hemispheric stroke: the European Cooperative Acute Stroke Study (ECASS). *JAMA* 1995;**274**:1017–25

9 The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. Tissue plasminogen activator for acute ischemic stroke. *N Engl J Med* 1995;**333**:1581–7

10 Hacke W, Kaste M, Fieschi C, *et al.* Randomized double-blind placebo-controlled trial of thrombolytic therapy with intravenous alteplase in acute ischaemic stroke (ECASS II). *Lancet* 1998;**352**: 1245–51

11 Bath P. Commentary: Alteplase not yet proven for acute ischaemic stroke. *Lancet* 1998;**352**:1238–9

12 European Study Group on Interferon beta-1b in Secondary Progressive MS. Placebo-controlled multicentre randomised trial of interferon beta-1b in treatment of secondary progressive multiple sclerosis. *Lancet* 1998;**352**:1491–7

Fetal distress

4212 delivery suite. With a rising tachycardia I jump out of my seat.
Could I pretend this hasn't happened and just not go?
My mind says YES. My conscience says NO.
'Hello it's Jo.' I answer my bleep.
'YOU the paed?' 'No I'm Jo,' I bravely repeat.
'Well get yourself down to room 8' says the voice.
I try to ask why, but she's gone, it's too late.
I pass a cleaner on my journey down that lonely corridor,
I plan a swift career change, I could handle emergency spillage on the floor.
Too soon I'm there. Deep breath Jo, and in you go.
'Here's the paediatrician.' I look over my shoulder.
Surely they're expecting someone wiser, someone older.
I scuttle past the screaming woman with her legs splayed in the air,
And seek refuge in the corner with my Resuscitaire.
A hundred bits of good advice run riot in my head
Including 'Use your common sense Jo', which is what my mother said.
With shaking hands I cut a tube and check the laryngoscope
'Please God don't let me have to use these.' Prayer is now my only hope.
I've never felt so sick with fear.
Suddenly the moment's here.
It's happening, it's not a dream.
And now I hear two people scream.
Baby 'crying and vigorous' Apgars 9 and 10.
Paediatrician passed out. In need of facial oxygen.
In a heap I lie and cautiously open one eye.
Another scarey midwife is staring down at me.
'Are you the paed? Get on your feet. Meconium room 3.'

Jo Frost
Southmead Hospital, Bristol BS10 5NB, UK