Improving statistical quality in published research: the clinical experience

Talk to be presented at "Statistical Methods for Pharmaceutical Research and Early Development", Lyon, France, September 27-29, 2010.

Abstract

Over the past 40 years, the quality of clinical research has improved greatly. I shall try to show how this has come about and identify key factors in this improvement. I shall go on to look at the position in non-clinical biomedical research and see whether there are any lessons to be drawn from the clinical experience.

Introduction

Thank you for inviting me to address this conference. It is a bit off my usual track, because for most of my career I have been engaged in clinical and epidemiological research. Not all, however. I started with three years in the agricultural chemical industry, but even there I was doing field trials, actually in muddy fields. Of course, this means that this conference is an opportunity for me to learn about something new. I intend to present a very personal account of how clinical research has changed in the 38 years since I started at St. Thomas' Hospital Medical School in 1972. I shall suggest some of the key factors in this change and tell you what I did to push things along. Then I shall try to compare the situation in non-clinical biomedical research and finally see whether I can make any constructive suggestions for the future.

Then and Now

When I began my medical research career, research published in elite medical journals was very different from now. I recently reviewed the *Lancet* and the *British Medical Journal* from September 1972, my first month. I was particularly interested in sample sizes for human studies and how these had changed (Bland 2009).

The *Lancet* contained 31 research reports which used individual subject data, excluding case reports and animal studies. The median sample size was 33 (quartiles 12 and 85). In the *British Medical Journal* in September 1972, there were 30 reports of the same type, with median sample size 37 (quartiles 12 and 158). I noticed that statistical considerations were almost entirely lacking from the methods sections of these papers.

For the 2009 paper, I compared I compared these papers to those in the same journals in September 2007, a similar 5 week month. For this talk, I have come up to date with July 2010, another 5 week month. In the *Lancet*, there were 16 such research reports, with median sample size 1626, IQR = 527 to 14774, two orders of magnitude greater than in 1972. In July 2010, the *BMJ* carried 15 such research reports, with median sample size 10170 (IQR 234 to 48649). Hence the median sample sizes have gone up from 33 and 37 to 1626 and 10170, two orders of magnitude greater. The sample size for studies in these journals has increased hugely.

I thought it would also be interesting to compare the methods of statistical inference employed. In 1972, statistical inference did not feature much in the abstract of papers. In 39 *Lancet* papers (including studies not on individual subjects), only five mentioned P values or significance in the abstract. In 32 *BMJ* papers, only four did

so. Many of these papers included statistical inference in the "Results" section of the paper. For the *Lancet*, 19 of the 39 quoted the results of significance tests, either as P values or test statistics, and one gave confidence intervals in graphical form (Pollack *et al* 1972). For the *BMJ*, 22 of the 32 papers gave the results of significance tests, none at all presented confidence intervals. Very little description of statistical methods appeared in "Methods" sections of these papers. Only three BMJ papers gave a reference for their statistical methods. One of the few that mentioned them at all (Bottiger and Carlson 1972) merely noted that 'Statistical analyses were performed using methods described by Snedecor (1956)', this being a standard statistical textbook, already superseded by the 1967 edition. This was also cited by Ellis (1972). Bishop *et al.* (1972) quoted Dixon and Massey (1951) a book then more than twenty years old.

In 2010, things were very different. In both journals, all papers included statistical inference in the abstract. For the *Lancet* 15 of the 16 papers had confidence intervals and 8 had P values, for the BMJ 13 of the 15 had confidence intervals, 7 had P values. So we have much greater sample sizes and much greater prominence for statistics in the papers. We also have a clear change of emphasis, from significance testing to estimation.

What Happened?

Several initiatives might have contributed to this change. They are not independent things, but different aspects of the same drive. Often it is hard to say exactly when these movements began, because a lot of people were involved in them.

Evidence-based medicine

Dave Sackett and Gordon Guyatt at McMaster University were leaders in the movement for "Evidence-based medicine". The earliest papers using this term began in the 1990s, but the ideas were around long before. The argument was that treatment decisions should be based on objective evidence rather than the evidence of experience and authority. Such evidence was going to include statistics. This was a doctor-led movement, but statisticians, as people whose business was the evaluation of evidence, were enthusiastic cheerleaders. Dave Sackett spent a sabbatical with us at St. Thomas's Hospital Medical School in the 1970s, so I was fortunate enough to know him while these ideas were forming.

Systematic review

An important aspect of evidence based medicine is systematic review, the idea that we should collect together all the trials which had been carried out of a therapy and try to form a conclusion about effectiveness. Iain Chalmers led a huge project to assemble all the trials ever done in obstetrics (Chalmers *et al*., 1989), a scheme that led to the even more grandiose Cochrane Collaboration, to do the same for all of medicine. This, too, was a doctor-led initiative, but statisticians were enthusiastic supporters, developing methods of data synthesis to combine the results of these trials where possible. Richard Peto springs to mind as very influential here. I recall him presenting in the early 80s a (never published) study of expert opinion on three approaches to the treatment of myocardial infarction, as expressed in leading articles in the *New England Journal of Medicine* and the *Lancet*, and contrasting this with the exactly opposite conclusions which he had drawn from a systematic review and fledgling meta-analysis of all published randomised trials in these areas.

Large simple trials

Richard Peto's favoured solution to the problem of inadequate sample sizes was large simple trials. Peto and Yusuf (1981) led the call for large, simple trials, the first being ISIS-1 (ISIS-1 Collaborative Group, 1986). This was spectacularly successful, as Peto *et al*. (1995) described. It probably explains the great increase in sample size reported from 1972 to the present. No clinical researcher with aspirations to be in the top flight can now be happy unless a trial with a four-figure sample size is in progress. I know that I have one!

Confidence intervals not P values

A very statistically led movement was to present inference using confidence intervals rather than significance tests. Gardner and Altman (1986) was a very important paper in this, which led to the *British Medical Journal* including this in its instructions for authors. Other journals, such as the Lancet, followed suit.

Quality assessments in journals

As Altman (1991) describes, there is a long history of articles criticising the quality of statistics in medical journals, but these mostly come from the mid-sixties onwards. Altman (1981) was an important article calling for improvement. These articles began to sting to journal editors into action and led to instructions to authors about statistical aspects of presentation of results.

Statistical referees

Another development following these reviews was the introduction of statistical referees for journals. By this I mean the systematic use of a panel of statisticians to referee all research papers before they appeared in the journal. The main difficulty with this is finding enough statisticians. Only major journals can manage to do it.

The CONSORT statement

CONSORT statement was first published in 1996 (Begg *et al.*, 1996). It has since been updated (Moher *et al.*, 2001) and produced several variations and imitators. This gave guidelines for reporting trials, encouraging researchers to provide information about methods of determining sample size, allocation to treatments, blinding, statistical analysis, etc. It has now been adopted by many journals as part of their instructions to authors.

Post Hoc Ergo Propter Hoc?

"After this therefore because of this" is a well established logical fallacy, often put as "correlation does not imply causation". We cannot know which, if any, of these forces is responsible for improvements in the statistical quality of the elite clinical literature.

My rôle in the campaign, or "What did you do in the war, Daddy?"

I have a very vivid memory of being at a meeting of teachers of statistics in medical schools, where a group of us held a discussion on a core curriculum for medical statistics. We reported our conclusions about t tests and chi-squared tests back, to have them demolished by David Clayton, who said that what he wanted students to learn was how to make estimates about the world and put confidence intervals around them. I saw that he was right and as soon as I got back to the office I redesigned my courses to put estimation first. From then on, in analyses carried out

for researchers I stressed confidence intervals. When my text book *An Introduction to Medical Statistics* (Bland 1987) first appeared, the chapter introducing confidence intervals came before that introducing significance tests, and their superiority was emphasised. Less praiseworthy was the blatant error in my explanation of what a confidence interval meant, but I fixed it in the second edition.

I wrote letters to journals when I saw blatant mistakes in statistical analysis. Sometimes the letters were published. The first one was actually the first publication of Bland and Altman (Bland and Altman 1997). Occasionally these mistakes were accepted by the authors, more often not, but they made the point to encourage future authors from copying flawed methods and interpretations.

Doug Altman and I wrote Statistics Notes in the *British Medical Journal*. These began in 1994 (Bland and Altman 1994) and continue sporadically ever since. We have published 55, with six other occasional authors and they had a mean of 115 citations by July 2010, a total of 6337.

I was involved in grant funding bodies and an ethics committee. On these I stressed the importance of correct statistical design and analysis. For example, I joined the Medical Research Council project board for health services and public health research. At my first meeting, there was a bid for a cluster-randomised trial, though the applicants did not use this term. In their sample size calculations and proposed statistical analysis they did not take any account of the clustering. This would mean that the trial would be underpowered and that any P values would be too liberal and confidence intervals too narrow. I explained this to the board and the proposal was rejected. At the next meeting, the same thing happened again. As meeting followed meeting, my colleagues started saying that they knew what I was going to say, but I might as well say it anyway. Later, this changed to "We know what you are going so say, don't bother!" Then a change occurred. Cluster randomised trials started being described as such and coming with estimates of intra-cluster correlation coefficients and proposals for multilevel modelling. I wondered what change had taken place in the world, without me knowing. Then I discovered that the MRC secretariat were warning applicants that cluster randomised trials which ignored the clustering would not get past the board --- "Professor Bland will stop it" --- and that they should find a statistician who understood these things. I think that was the most effective piece of statistical education that I ever did.

Is it all over?

Although it would be nice to report that clinical research is now statistically flawless, this is not so. Things are much better in the major journals. In the specialist clinical journals, where statisticians seldom venture, things can go on much as before.

An example is given by the Boots "anti-aging" cream trial, published last year. This trial received wide media publicity as the first anti-aging cream proven to work in a randomised controlled clinical trial. It was published in the British Journal of Dermatology (Watson *et al.* 2009). I read the paper and found that 60 volunteers were randomised in groups of 30 to either the 'anti-aging' product or the vehicle without the active ingredient. The authors reported that after six months 43% of participants receiving the 'anti-aging' cream had improved appearance of wrinkles, compared to 22% of those receiving the placebo and this was what was picked up by the media.

The authors report four outcome measures: fine lines and wrinkles, dyspigmentation, overall clinical grade of photoageing, and tactile roughness, each measured on a scale of 0 to 8 at baseline, 1, 6, and 12 months. There was no mention that any of them being a prespecified primary outcome, so we might surmise that a significant difference in any variable would be taken to indicate evidence of a treatment effect. The trial is entirely analysed in terms of P values, so prudence should lead us to adjust for multiple testing. The easiest way to do this for a published paper is to use the Bonferroni correction. The authors did not do this, but if they had done we would have conventional significance if for any of the four outcomes the P value multiplied by 4 is then less than 0.05. If we were to include the 6 and 12 months results in the same analysis, we would multiply by 8. There was also a measure of fibrillin-1 in 28 of the subjects, measured on a 5-point scale in a biopsy at six months, which could also be included to give 5 tests at six months or 9 tests overall.

For wrinkles at six months, the authors gave the results of significance tests comparing the score with baseline for each group separately, reporting the active treatment group to have a significant difference and the vehicle group not. This is a classic statistical mistake. The difference within a group being not significant does not imply that there is no difference in the population or tell us much about the size of any difference that might exist. We should compare the two groups directly. The paper does include some data for the improvement in each group, 43% for the active group and 22% for controls, as picked up by the media. No P value is given, but in the discussion the authors acknowledge that this difference was not significant.

The *British Journal of Dermatology* published my letter (Bland 2009b) and a different version subsequently appeared in *Significance* (Bland 2009c). This happened, of course, only because the publicity generated by Boots brought the paper to my attention.

Non-clinical biomedical research

I have remarked several times that laboratory research is the next area for statisticians to become involved. My very limited observation of this literature is that in it research scientists do their own statistics and often do them badly.

Here is an example which I have been inflicting on my students for the past few years. Temme *et al.* (2001) compared two genetic strains of mice, wild-type and connexin32-deficient. They measured the diameters of bile canaliculi in the livers of wild-type and C02-deficient animals, making several observations on each liver.

Their results are shown in the Figure.

I think there is a fairly obvious problem with the units of analysis here. They have two groups with three observations in each, not groups of 280 and 162 observations. Hence their significance test is quite wrong.

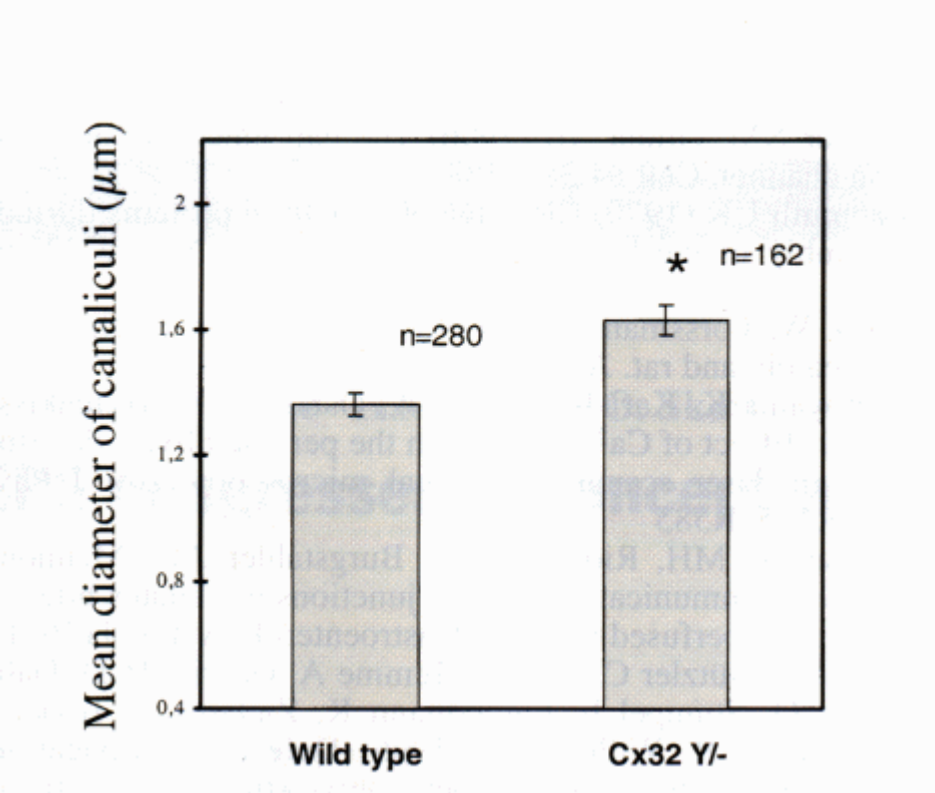

Figure. Morphometric analysis of the diameter of bile canaliculi in wild-type and C02-deficient liver. Means±SEM from three livers. *P<0.005, after Temme *et al.* (2001)

To save me doing a review of the laboratory literature, Kilkenny *et al.* (2009) carried out a review of reporting, experimental design and statistical analysis in published biomedical research using laboratory animals. They analysed 271 publications and reported that in only 59% the hypothesis or objective of the study and the number and characteristics of the animals used were reported. Most of the papers surveyed did not use randomisation (87%) or blinding (86%), to reduce bias in animal selection and outcome assessment. Only 70% of the publications that used statistical methods described their methods and presented the results with a measure of error or variability.

What next?

Our best allies are journal editors. Once they are convinced that there is a serious problem, they usually want to do something about it.

Reviews of statistics used in particular journals are a good starting point. They are quite easy to do, best done more by than one statistician independently. They give a statistical publication, too, which is always useful for the biomedical statistician. Emulating reviews of clinical journals, Jeremy Miles reviewed two psychological journals and found two instances of “P<0.0” (Miles and Hempel, 2005). I’ll repeat that, yes, TWO instances of “P<0.0”, showing that the authors just did not understand what they were doing, because, of course, P values cannot be negative.

Case studies of examples where wrong conclusions have been drawn as a result of statistical mistakes provide very powerful evidence, if you can find them. Richard Peto's review of myocardial infarction is a good example.

When you do see mistakes in published research, write a letter to the journal. Harry them!

Finally. be positive. We want to help. Try offering statistics articles. I think a few on the benefits of randomisation and blinding would be good starting point.

References

Altman DG. (1981) Statistics and ethics in medical-research. 8. Improving the quality of statistics in medical journals. *British Medical Journal* **282**: 44-47.

Altman DG. (1991) Statistics in medical journals - developments in the 1980s. *Statistics in Medicine* **10**: 1897-1913.

Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF. (1996) Improving the quality of reporting of randomized controlled trials - The CONSORT statement. *JAMA-Journal of the American Medical Association* **276**, 637-639.

Bishop MC, Woods CG, Oliver DO, Ledingham JGG, Smith R, Tibbutt DA. (1972) Effects of haemodialysis on bone in chronic renal failure. *British Medical Journal* **3**: 664-667.

Bland JM. (2009) The tyranny of power: is there a better way to calculate sample size? *British Medical Journal* **339**: b3985.

Bland JM. (2009b) Evidence for an 'anti-ageing' product may not be so clear as it appears. *British Journal of Dermatology* **161**, pp1207–1208.

Bland M. (1987) *An Introduction to Medical Statistics*. Oxford University Press, Oxford.

Bland M. (2009c) Keep young and beautiful: evidence for an "anti-aging" product? *Significance* **6**, 182-183.

Bland JM, Altman DG. (1977) Enteric disease in San Francisco. *Lancet* 2, 306.

Bland JM, Altman DG. (1994) Statistics Notes. Correlation, regression and repeated data. *British Medical Journal*, 308, 896.

Bottiger LE, Carlson LA . Relation between serum-cholesterol and triglyceride concentration and hemoglobin values in non-anemic healthy persons. *British Medical Journal* 1972; **3**: 731-3.

Chalmers I, Enkin M, Keirse MJNC. (eds) *Effective Cure in Pregnancy and Childbirth*, Oxford University Press, Oxford, 1989.

Dixon WJ and Massey FJ. (1951) *Introduction to Statistical Analysis* New York: McGraw Hill.

Ellis FR, Keaney NP, Harriman DGF, Sumner DW, Kyei-Mensah K, Tyrrell JH, Hargreaves JB, Parikh RK, Mulrooney PL (1972) Screening for malignant hyperpyrexia. *British Medical Journal* **3**: 559-561.

Gardner MJ and Altman DG. (1986) Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal 292*: 746-50.

Goodman SN, Altman DG, George SL. (1998) Statistical reviewing policies of medical journals - Caveat lector? *Journal of General Internal Medicine* **13**: 753-756.

Guyatt G. Evidence-based medicine - a new approach to teaching the practice of medicine. *Jama-Journal Of The American Medical Association* 1992; **268**: 2420-2425.

ISIS-1 (First International Study of Infarct Survival) Collaborative Group. (1986) Randomized trial of intravenous atenolol among 16,027 cases of suspected acute myocardial infarction. ISIS-I. *Lancet* ii: 57-66.

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, MFW., Cuthill, IC., Fry, D., Hutton, J., Altman, DG. (2009) Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS ONE* 4(11), e7824.

Miles JNV, Hempel S. (2005) The presentation of statistics in clinical and health psychology research. In: *Proceedings of the British Psychological Society*, **13**, 185.

Moher D, Schulz KF, Altman DG, CONSORT Group. (2001) The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* **357**: 1191-4.

Peto R, Collins R, Gray R. (1995) Large-scale randomized evidence: large, simple trials and overviews of trials. *Journal of Clinical Epidemiology* **48**: 23-40.

Peto R, Yusuf S. (1981) Need for large (but simple) trials. *Thrombosis and Haemostasis* **46**: 325-325.

Pollack M, Nieman RE, Reinhard.JA, Charache P, Jett MP, Hardy PH. (1972) Factors influencing colonisation and antibiotic-resistance patterns of gram-negative bacteria in hospital patients. *Lancet* **2**: 668-1.

Snedecor, G.W. (1956). *Statistical Methods*. Ames, Iowa, Iowa State College.

Temme A, Stumpel F, Rieber GSEP, Willecke KJK, Ott T. (2001) Dilated bile canaliculi and attenuated decrease of nerve-dependent bile secretion in connexin32-deficient mouse liver. *Eur J Physiol* **442**, 961-966.

Watson REB, Ogden S, Cotterell LF, Bowden JJ, Bastrilles JY, Long SP, Griffiths CEM. A cosmetic 'anti-ageing' product improves photoaged skin: a double-blind, randomized controlled trial. *British Journal of Dermatology* 2009: DOI 10.1111/j.1365-2133.2009.09216.x