A Historical Reminiscence

Martin Bland

Emeritus Professor of Health Statistics
University of York

Talk presented at the celebration of twenty years of the Centre for Statistics in Medicine, 29th March 2017

This month I reached my 70th birthday and this led me to reflect that history was no longer just something I read about, but also something which I had experienced.

Old enough to remember rationing, the Suez crisis, the Summer of Love, to have seen Pink Floyd play a Saturday night dance at Imperial College. Older than the oldest digital computer, I have used paper tape punch, card punch, and line printer. Born before the publication of what is usually regarded as the first randomised control clinical trial. I am going to think today particularly about the history of medical statistics and even more particularly about the part in it played by myself and by Doug Altman.

My story begins in September 1972, when I arrived at the Department of Clinical Epidemiology and Social Medicine at St Thomas's Hospital Medical School, where Doug had been in post for about two years. I had spent my first three statistical years in agriculture with ICI Plant Protection, following four years at Imperial College. There I had the privilege of being taught by Sir David Cox, who introduced me to the idea of variance one Tuesday morning in October 1965. Before I applied to join Tommy's, I had to look up "epidemiology" in the dictionary; it was not then a word in popular use. Tommy's was unusual in having a moderately large group of statisticians: six academics and several juniors. In London, only LSHTM had so many. When I moved to St. George's four years later, I was their first.

Medical research then was very different to now. I reviewed *The Lancet* and *BMJ* for September 1972, the week I began my career in medical statistics. Samples were small. In 61 research reports which used individual subject data, excluding case reports and animal studies, the median sample size was 36 (quartiles 12 and 86). In all 71 research reports, excluding case reports, the methods of statistical inference employed, if any, were overwhelmingly significance tests. In the Abstracts of the 71 papers, nine mentioned P values or significance, none mentioned confidence intervals. In the "Results" section of the papers, 41 of 71 papers 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).

Very little description of statistical methods appeared in "Methods" sections of the papers. Three papers gave a reference for their statistical methods. One of these merely noted that "Statistical analyses were performed using methods described by Snedecor (1956)" (Bottiger and Carlson 1972), a standard statistical textbook, already superseded by the 1967 edition.

That was clinical research, epidemiology was methodologically much more advanced. I joined ongoing studies of smoking and respiratory disease in thousands of schoolchildren. Doug was working on the even bigger national study of health and growth. In clinical research, there were still arguments about whether randomisation was either ethical or desirable.

But change was on the way. Evidence-based medicine was in the air. This was a doctor-led movement, two of the principal movers being Gordon Guyatt and David Sackett at McMaster, Ontario. The earliest reference I can find to this term was a paper by Guyatt in JAMA in 1992, (Guyatt 1992) but the ideas were being developed much earlier. In the early 1970s, David Sacket visited our department at Tommy's for a sabbatical, where he was a big influence on both Doug and myself. I think we both learnt a lot from him. Of course, statisticians, including ourselves, were enthusiastic supporters of evidence-based medicine. Our business was evidence.

An important component of evidence-based medicine was the promotion of systematic reviews. 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). Now, the Cochrane Collaboration, which he founded, aims to do the same for all of medicine. I think this was another doctor-led initiative. Again, statisticians were enthusiastic supporters, developing methods of data synthesis to combine the results of these trials where possible. At first I was sceptical, thinking that we should do the trials well enough in the first place, so that another trial would be both unnecessary and unethical. However, I realised that such counsels of perfection were pointless and this was not going to happen. I started both to do meta-analyses and to teach others how to do them.

However, the desire for definitive trials was not just mine and Richard Peto led the call for large, simple trials (Peto and Yusuf 1981). The first such trial was ISIS-1. The paper in the Lancet was titled "Randomized trial of intravenous atenolol among 16,027 cases of suspected acute myocardial infarction" (ISIS-1 group 1986), the only paper I knew with the sample size in the title. Now, of course, big trials are everywhere and samples of thousands are no longer the preserve of the epidemiologists.

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. Eventually, even the *New England Journal of Medicine* did so.

Statistical reviews of journals. 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,1991). Altman (1981) was an important article calling for improvement. These articles began to sting journal editors into action and led to instructions to authors about statistical aspects of presentation of results.

Statistical refereeing. Following reviews of statistics, journals began to introduce statistical referees. Journals recruited statisticians to referee all research papers before they appeared in the journal. The main difficulty is finding enough statisticians. With the proliferation of journals, it seems unlikely that journals other than the majors could do this.

Anyone who has done a few systematic reviews will recognise that a big problem can be trying to work out what researchers actually did and whet the precise results of a study are. Sometimes we might know how many participants were recruited, but not how many were observed at outcome. We might know that a difference was not significant, but not how big it was, or what the standard deviation of the measurement was. This led to the development of the CONSORT statement. First published in 1996 (Begg *et al.*, 1996), this provides guidelines for reporting trials, encouraging researchers to provide information about methods of determining sample size, allocation to treatments, blinding, statistical analysis, precise results, etc. This has since been updated (Moher *et al.*, 2001) and continues to be improved (http://www.consort-statement.org/consort-statement/), and has produced several variations and imitators. It has now been adopted by many journals as part of their instructions to authors. Doug was asked to referee the original CONSORT statement, and was then asked to join the group for subsequent iterations.

I think that another important influence on research quality is the writing of critical letters to journals, pointing out mistakes in research methods and reporting. Doug and I made our first unsuccessful attempt at a joint publication with a letter to Lancet. This pointed out that a paper describing allocation as "more or less randomly" was misleading. We wrote "we wish to point out that allocation is either random or it is not – there is no intermediate state. Nearly random is not random." The Lancet suggested that we contact the authors, which was hardly the point. I was pleased to see that, some time later, Richard Peto mentioned this unfortunate phrase in a review; good for him.

Our first actual publication does not appear on either of our official CVs: a jokey letter in *What's Brewing?*, the magazine of the Campaign for Real Ale. This responded to a report on the possible adverse health effects of drinking keg beer. We wrote:

> It is hard to believe that the stuff can do the drinker any good, but to prove that it actually does him harm, surveys and experiments must be designed most carefully to forestall the inevitable accusations of bias. As statisticians working medical research and good CAMRA men, we would be keen to help in assembling data or designing and analysing controlled experiments — provided enough volunteers can be found to swallow the suspect brew! — Martin Bland & Douglas Altman, London SE11. (Bland M and Altman D 1975)

We also collaborated on one other literary effort at the time, an allegedly humorous sheet distributed at the Christmas party in our Sancroft Street office, entitled *The Sanitary Crofter*. It will give you a flavour of this effort, like the *What's Brewing* letter composed largely in the pub, if I tell you its five authors signed themselves The Four Degrees of Freedom.

It was not until we had both left STHMS that we had our first joint academic publication, the first of 100. This was a letter in the *Lancet*, criticising a study of enteric disease in San Francisco (Bland and Altman 1977). Doug was at MRC Northwick Park, I was at St. George's Hospital Medical School. As I was the only statistician there, naturally I wanted contact with others in my field and I spoke to Doug often.

This was also how we discovered that we had both come across the correlation and agreement problem. David Robson, one of my clinical colleagues, handed me a paper, saying that there was something wrong with it, but he didn't know what. The authors, Keim *et al.* (1976) were looking at the agreement between two methods of measuring cardiac stroke volume. They had pairs of measurements, one by each method, plotted a scatter diagram and calculated a correlation coefficient and associated P value. They also make a series of pairs of measurements on 20 individuals and for each patient they calculated the correlation. They found that only one of the 20 had a significant correlation and concluded that their methods did not agree. I realised that unless the actual stroke volume changed during this process, they were just correlating measurement error and their result was inevitable no matter how good the agreement was. I mentioned this to Doug and he said that he had come across a very similar thing with blood pressure. The correlation was always greater with systolic pressure than with diastolic (Hunyor *et al.* 1978, Laughlin *et al.* 1980). Of course, this is because systolic pressure has a greater variance than diastolic. I hand wrote a rough draft of the start of a discussion. Doug searched for other examples and came up with two different misleading approaches. The Institute of Statisticians announced a conference on health statistics. We decided to put in our first statistical abstract.

We thought that standing up, saying everyone was doing it wrong, then sitting down might fall a bit flat. So we brainstormed what a statistician would do and it took us half an hour or so to come up with the limits of agreement method. We submitted the abstract and gave the talk in Cambridge in 1981.

Altman and Bland in Cambridge, 1982

We thought the method was so obvious that we did not claim originality. I was convinced that someone would say something along the lines of "Fisher did that in 1937", but they didn't. So we drafted a paper for publication. We were both so nervous that Doug sent the draft to Michael Healey and I sent it to Sir David Cox! They were both very generous and

eventually we submitted it to the late lamented *Statistician* and it was published (Altman and Bland 1986). Our names were in alphabetical order. David Robson, who drew the problem to my attention in the first place, received only an acknowledgement.

Our friends liked the paper, but elsewhere researchers carried on correlating. People suggested that we write a version for a clinical readership, with a worked example. So I decided to collect my own data, comparing a mini Wright peak flow meter with a standard meter. I collected two measurement of PEFR by each method, in random order, from ourselves, colleagues, and my family. This is one of the few published data sets including the author's parents and parents-in-law. Then we wrote the paper. Where to send it? "Why not the *Lancet*?", I said, they will give us a quick reply and we might get some helpful comments when they turn it down. So off it went, name order reversed so I was first. The week before Christmas, 1985, I was phoned by David Sharpe, the deputy editor of the *Lancet*. He said they would like to publish it. But it was too long. He heard my sigh down the phone and said the *Lancet* would do the editing, rather than us! After Christmas it came back. It was definitely better, but something vital, so I thought, was missing. I asked David Sharpe if we could have one paragraph back, the confidence interval for the limits. "Is that really important?" he asked. I assured him that it was! So back it went and the paper was published (Bland and Altman 1986).

To our amazement, the *Lancet* paper was a staggering success. It led to the *Statistician* paper being cited, too. In 1992, to my amazement, the Institute for Scientific Information asked us for commentary on the two papers as a citation classic for *Current Contents* (Bland and Altman 1992), only six years after the *Lancet* paper was published. It became the most highly cited paper in the *Lancet*, one of the highest cited statistical papers, and in 2014 was reported be one of the top 30 most highly cited papers in any field (Van Noorden *et al*. 2014). This doesn't mean we are top scientists; Einstein isn't in the list, nor Watson and Crick, nor R.A. Fisher. We were asked many questions about how to calculate limits of agreement in more complex situations, which we tried to answer, and we published more papers on the topic. We now have four publications on agreement with more than 1,000 citations on WoS (Altman and Bland 1983, Bland and Altman 1986, Bland and Altman 1995, Bland and Altman 1999) and five which are the most highly cited in their journals (Altman and Bland 1983, Bland and Altman 1986, Bland and Altman 1999, Bland and Altman 2003, Bland and Altman 2007).

Doug moved to ICRF, Lincoln's Inn Fields. I was due a sabbatical. I had two small children, I didn't want to go away, so I spent it with Doug. We wrote and submitted an RSS Ordinary Meeting paper, published 1991 (Altman and Bland 1991). Not one of our greatest hits, only 73 citations, but a great experience.

In 1994 Doug was invited by BMJ to contribute a series of fillers to be called Statistics Notes. Paper BMJ needed to avoid gaps on the page. He, very sensibly, said that to keep him from error he would need a co-author and I was the person for the job. The BMJ liked that and so our series began (Bland and Altman 1994), now extended to 65 with eight co-authors, but always one of us. We hoped that people would read them, but thought that nobody would

remember them and that they would lengthen our CVs, but no more. Then people began to cite them and one, Bonferroni (Bland and Altman 1995), passed 1,000 citations. Now it has been joined by Cronbach's alpha (Altman and Bland 1997). Total citations for the Notes reached 15,332 on March 7th 2017.

Meanwhile, Doug had left London for Oxford and the formation of the Statistics Unit. We continued to work closely, because the technological revolution which has so influenced statistical practice had also brought about email. Soon we were in frequent contact, exchanging drafts. We still met often in London at RSS meetings and, because the rail link between London and Oxford is good, it was easy for me to travel out here to meet. Then, in 2003, I too left the big city, for historic but small York. We continued to write together, but less frequently as time took its toll and the slow rail link between York and Oxford made meeting infrequent. But we go on and already have a minor publication accepted in 2017 (Altman and Bland 2017).

In Oxford, Doug's career moved in new directions. First, he was leading the Unit, which must be a time-consuming role, becoming more so as the unit grew. I have tried to avoid it; when I moved to York it was on condition that I did not have to be head of or chairman of anything. Second, he was seeing other collaborators, developing an ever-widening circle of colleagues from around the world. This led to involvement with many others interested in the quality of research reporting, including CONSORT, which he joined after acting as a referee for the first iteration, and the EQUATOR Network, which he leads.

Did it all work? Is medical research of higher quality now? I repeated my review of Lancet and BMJ for Feb 2017.

Sample size: then, 61 papers, median = 36, IQR = 12 to 86, now, 25 papers, median = 4,383, IQR = 500 to 21,395.

P values or significance in Abstract: then, 9/71, now, 18/30 papers.

P values or significance in paper: then, 41/71, now, 24/30

Confidence intervals in Abstract: then 0/71, now 27/30

Confidence intervals in paper: then 1/71, now 27/30

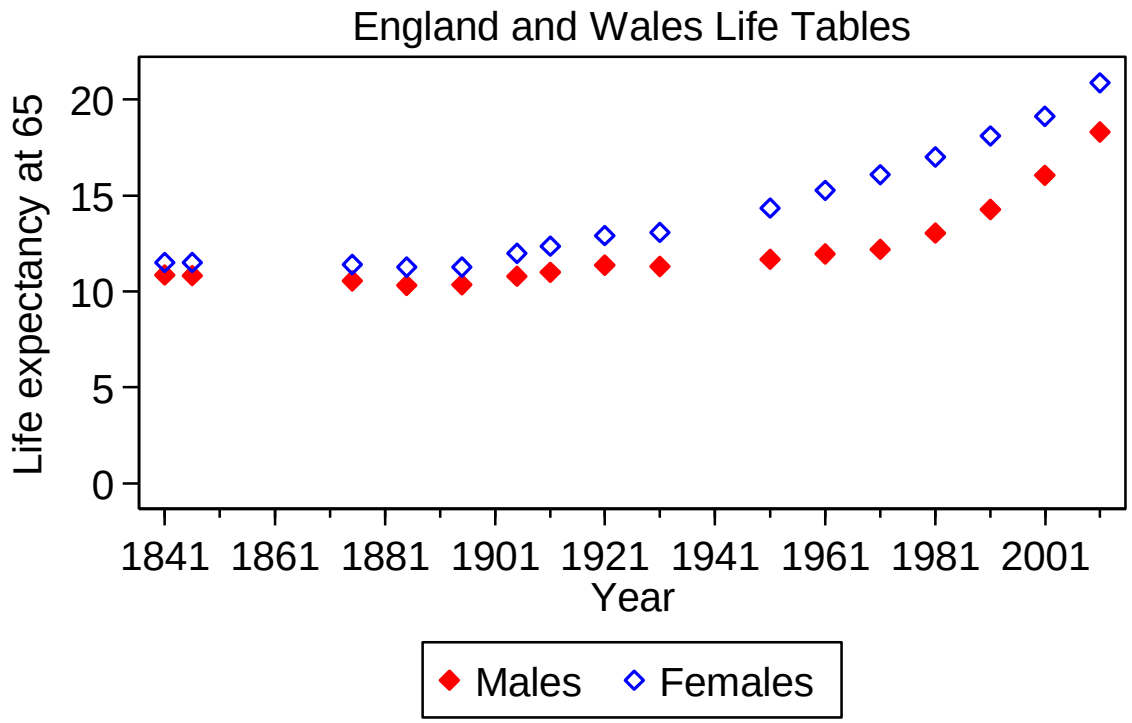
No statistical inference: then 30/71, now 3/30

Methods sub-section: then 0/71, now 28/30

Methods paragraph or reference only: then 3/71, now = 2/30.

So the quality of research, at least as reported in the top medical journals, has improved greatly, with far bigger samples, far better inference. But has this had an effect on the quality of medicine itself? Are we living longer, healthier lives? I think that I am! But how can we find an index for the population? I thought of expectation of life. Not expectation of life at birth, which depends on things like economic development, housing, nutrition, etc. I chose

life expectancy at age 65, as this age group is a major consumer of health care and I knew that before 1971, life expectancy at 65, certainly for men, had changed little since the 19th century. This is what I found:



Expectation of life at 65 rose for women after 1900, much more than for men. It has been suggested that this is because of the fall in pregnancies per woman leading to women reaching age 65 fitter and healthier than during the nineteenth century. For men it stays almost flat until 1981. It then rises faster than for women and men have almost caught up. Women living longer than men was a twentieth century phenomenon, which may be coming to its end.

Why this happened if difficult to say. Did all the things I described cause this? Post hoc ergo propter hoc, after this therefore because of this, is a well known logical fallacy. As statisticians say, correlation does not imply causation. The decline in cigarette smoking undoubtedly was a big factor, but that arose from great epidemiology from Doll and Hill (1950, 1956). I think medicine is better than it was and I think better medical research is the reason for this, though I cannot prove it. But need I say that the first Bland and Altman publication was in 1977? Post hoc ergo propter hoc!

Meanwhile, the wider medical literature can still be a mess, there is much still to do. The old guard will step back. Over to you.

Let us make all evidence trustworthy!

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.

Altman DG, Bland JM. (1983) Measurement in medicine: the analysis of method comparison studies. *The Statistician* 32, 307-317.

Altman DG, Bland JM. (1991) Improving doctors' understanding of statistics (with discussion and reply). *Journal of the Royal Statistical Society, Series A* 154, 223-267.

Altman DG, Bland JM. (1997) Statistics Notes. Cronbach's alpha. *British Medical Journal* 314, 572.

Altman DG, Bland JM. (2017) Assessing agreement between methods of measurement. *Clinical Chemistry,* in press.

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.

Bland M and Altman D. (1975) Health damage. *What's Brewing* March 1975, p2.

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

Bland JM, Altman DG. (1995) Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet* 346, 1085-7.

Bland JM, Altman DG. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* i, 307-310.

Bland JM, Altman DG. (1992) This week's citation classic: Comparing methods of clinical measurement. *Current Contents,* CM20(40) Oct 5, 8.

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

Bland JM, Altman DG. (1995) Statistics Notes. Multiple significance tests: the Bonferroni method. *British Medical Journal* 310, 170.

Bland JM, Altman DG. (1999) Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 8, 135-160.

Bland JM, Altman DG. (2003) Applying the right statistics: analyses of measurement studies. *Ultrasound in Obstetrics & Gynecology* 22, 85-93.

Bland JM, Altman DG. (2007) Agreement between methods of measurement with multiple observations per individual. *Journal of Biopharmaceutical Statistics* **17**, 571 – 582.

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.

Doll, R. and Hill, A.B. (1950). Smoking and carcinoma of the lung. *British Medical Journal*, **ii**, 739–48.

Doll, R. and Hill, A.B. (1956). Lung cancer and other causes of death in relation to smoking: a second report on the mortality of British doctors. *British Medical Journal*, **ii**, 1071–81.

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

Guyatt G. (1992). Evidence-based medicine - a new approach to teaching the practice of medicine. *JAMA* **268**: 2420-5.

Hunyor, S. M., Flynn, J. M. and Cochineas, C. (1978). Comparison of performance of various sphygmomanometers with intra-arterial blood-pressure readings. *British Medical Journal 2*, 159-62.

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-l. *Lancet* ii: 57-66.

Keim, H. J., Wallace, J. M., Thurston, H., Case, D. B., Drayer, J. I. M. and Laragh, J. H. (1976). Impedance cardiography for determination of stroke index. *Journal of Applied Physiology 41*, 797-9.

Laughlin, K. D., Sherrard, D. J. and Fisher, L. (1980). Comparison of clinic and home blood-pressure levels in essential hypertension and variables associated with clinic-home differences. *Journal of Chronic Diseases 33*, 197-206.

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, 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.

Van Noorden R, Maher B, Nuzzonoorden R. (2014) The top 100 papers. *Nature* **514**: 550.