The evolution of ways of deciding when clinical trials should stop recruiting

Peter Armitage
Emeritus Professor of Applied Statistics, University of Oxford, 2 Reading Road, Wallingford, OX 10 6DP, UK
Corresponding author: Peter Armitage. Email: armitage55@btinternet.com

[Based on an interview of Peter Armitage (PA) by Iain Chalmers (IC) on 9 September 2013, in Wallingford, Oxfordshire]

IC: *You have spent more than half a century thinking about ways of deciding when clinical trials should stop recruiting. I would be surprised if there is anyone else in the world who has comparable experience. I am very grateful to you for being willing to be interviewed about the ways your views have evolved over that time.*

I have a memory of reading in one of Austin Bradford Hill's articles on 'the clinical trial' that deciding when to discontinue recruitment to a trial often presents a quandary. Assuming that I am remembering his view correctly, do you share it? Can you recall where he wrote it?

PA: I don't know of any specific quotation from Bradford Hill's writings, but I am sure he would have taken that view, which is certainly true. Curiously, in his expository papers on clinical trials he does not seem to spend much time on matters of trial size, being understandably more concerned about bias in assignment and assessment.

The basic quandary is as follows. If the data, however imprecisely, suggest that there is a difference between treatments, the trial may be stopped too early and lead to an imprecise, inconclusive result. Despite the resulting uncertainty, it may be difficult to arrange further trials addressing the same question because of ethical concerns about further use of an apparently poorer treatment. On the other hand, if a trial goes on 'too long' it may have allowed too many patients to be treated with an inferior regimen.

IC: *I believe your interest in ways of deciding when trials should stop recruiting originated in statistical approaches that you had been using in industry. Is that correct? Did the mathematical paper you published in the* Journal of the Royal Statistical Society *in 1950 relate to your work in industry?*

PA: Yes. During the war I worked in a Ministry of Supply unit concerned with industrial sampling inspection and quality control, set up as part of the major push on armaments production. I was in the sampling inspection research group (SR17) led by G.A. Barnard. Typical products, such as fuses, were produced in large batches which were inspected by sampling, for example, by taking, say, 30 fuses and classifying them as defective or not. The batch would be failed if there were too many defectives and passed if there were very few. There was a clear advantage in taking an initial small sample and giving a pass/fail verdict if the answer was clear, and adding one or more additional samples in more equivocal cases. The sample size thus depended on the data. Work went on in the UK and USA on variants of this idea, leading to more general strategies of sequential sampling where the progression to larger samples was more continuous, with possible stopping at many stages. The theory was generalized by Abraham Wald, in a report which was sent to us in confidence and the basis of his 1947 book.[1] I worked on various extensions of Wald's methods, some of which were published later.

There was an analogy here with clinical trials, except that if a clear difference in effectiveness between treatments appears early this may lead to early termination on ethical, rather than economic, grounds.

After a final year back at Cambridge in 1946–1947 I was signed up for a permanent post in the scientific civil service, at the National Physical Laboratory, Teddington. I knew virtually nothing about medical statistics and was surprised and pleased to be offered a post under Austin Bradford Hill (ABH) in the Medical Research Council's Statistical Research Unit at the London School of Hygiene and Tropical Medicine, starting December 1947. This came about because Edgar Fieller, my boss at the National Physical Laboratory, and Donald Reid (ABH's head of epidemiology) commuted to London together from their Surrey suburb, and

Reid asked Fieller if he had a suitable young man on offer!

This introduced me to medical statistics, but I retained an interest in sequential analysis. My 1950 paper in the *Journal of the Royal Statistical Society*[2] was not directly about clinical trials, but it contributed to my later thoughts about introducing sequential analysis in clinical research. ABH encouraged this, and asked me to write a report to show how sequential analysis might have been used in trials already completed, using the original trial data. The report, now lost, showed that sequential plans with the same power as the actual trials would have reached the same conclusions with savings in trial numbers; this would tend to happen if the original trials showed a statistically significant difference between treatments, and this is the situation when the ethical case becomes strong. This report led me to write the paper that was published in the *Quarterly Journal of Medicine*,[3] of which more later.

IC: *Please outline for non-statisticians what sequential methods of trial analysis are.*

PA: The general idea of sequential analysis in clinical trials is to have a plan that allows results to be accumulated and analysed continuously, often conveniently by plotting on a chart. The simplest case would be a trial to compare two treatments, giving a sequence of 'preferences' for one or other of the treatments. These might be obtained by pairing patients, randomly allocating them to the two treatments, and finally giving a preference to whichever does better. Or, in a crossover trial, a patient may be given each treatment on different occasions, in random order, with a preference given to the treatment with the better outcome. The plan would control error probabilities, that is, the 'Type 1' error of claiming a statistically significant difference when the treatments are really equally effective, and the statistical 'power' – the chance of detecting a real difference if it is present. The plans would ensure that big effects are likely to be detected quickly. Later designs by me and many others introduced developments which enabled the individual responses to treatment to be measured more subtly than by mere preferences, for example, by measuring, say, change in lesion size, or by the time taken to reach some critical event, say, duration of symptom remission.

Why have a special theory for sequential analyses? If accumulating data are analysed continuously, the usual formulae for estimating error probabilities (which apply to single analyses) are not valid. If you continually test for statistically significant differences between treatments you run a higher chance of finding one, purely by chance, and risk stopping the trial with claims of a breakthrough which are not justified by the data.

IC: *The earliest example of this approach being used in clinical trials in the* James Lind Library *was reported by Newton and Tanner.[4] Did these researchers seek your advice? Are you aware of any earlier examples?*

PA: The first major literature reference was the description by Bross[5] of two specific plans (rather than a general theory), but he did not illustrate these plans with data from actual trials. Newton and Tanner[4] followed one of Bross's plans. I think they did this before meeting me. I don't know of earlier examples. Most of the trials I advised on were at the end of the 1950s or later.

IC: *In 1953 you submitted a very substantial paper to the* Quarterly Journal of Medicine. *When it was published in 1954, it appears to have been the first detailed (19-page) exploration of the applicability of sequential methods in medicine. It is a very technical paper yet you submitted it to and it was accepted by a medical journal. Tell us about the pre-publication and post-publication history of the article.[3]*

PA: I have mentioned the (lost) internal report that ABH asked me to prepare. He thought the idea of applying sequential methods in clinical trials was worth exploring and I suspect he raised this with one of the editors of the *Quarterly Journal of Medicine*. At any rate I was given a good welcome by the Journal and enabled to formulate my general ideas for sequential analysis in clinical trials.

IC: *At least four clinical trials using sequential analysis were published in the British literature between 1956 and 1959, three of them specifying the method in the titles.[4,6,7,8]*

Were these methods also used outside the UK? If so, where?

PA: I don't know of use outside the UK during this period. The last three came from approaches to me or the MRC Statistical Research Unit by people who had read about the idea. They all used a form called 'restricted' sequential designs[9] which, like Bross's two examples, were 'closed', in that an upper limit was declared to the number of preferences to be recorded. A number of other trials I helped with in the 1960s followed similar methods. I particularly enjoyed a series of trials in Nigeria and India on tetanus antitoxin, beginning with Brown et al.[10] showing that antitoxin worked but that a large dose was apparently no better than a small one.

IC: *I suppose interest in these methods must have been substantial because you prepared a book on* Sequential Medical Trials, *which was published in*

1960.[11] *Did you approach a publisher with a suggestion for the book?*

PA: After a year in the USA (1957–1958) and talking to people there, I thought there was room for a short book. I approached the medical publisher HK Lewis with the first chapter, but they were not interested. I then tried Blackwell Scientific Publications and had an enthusiastic welcome, especially from Per Saugman, the leading light there. They were equally enthusiastic about the second edition.[12] But publication of that second edition really marked the end of my active engagement in this area, especially as I left my post at the London School of Hygiene and Tropical Medicine in 1976 for a chair at Oxford which was not specifically medical.

IC: *How was the book generally received? Were there any obvious differences in its reception by statistician reviewers and medical reviewers?*

PA: It was received politely and on the whole favourably, I think – by both camps! In some ways it must have been an awkward book to review, being too non-mathematical for academic statisticians but perhaps a headache for non-statistical physicians.

IC: *One statistician reviewer – Frank Anscombe*[13] *– claimed that 'Sequential analysis is a hoax' and that 'The experimenter should feel entirely uninhibited about continuing or discontinuing his trial, changing his mind about the stopping rule in the middle, etc., because the interpretation of the observations will be based on what was observed, and not what might have been observed but wasn't.'*

In your response to Anscombe – a Bayesian – you noted (inter alia) that you were (i) 'not convinced that the interests of scientific communication would be served by encouraging the research worker to express his usually vague prior beliefs in quantitative terms'; (ii) 'that trials on the scale envisaged by Anscombe's theory seem beyond the reach of present resources'; and (iii) by the time that 'a difference of 3 or 4 times its standard error' had been reached 'the pressure to stop the trial would be overwhelming'.[14] *Were Anscombe's views a foretaste of the subsequent failure of Bayesian approaches to be adopted by clinical trialists? And was your reference to 'a difference of 3 or 4 times its standard error' a foretaste of the stopping guidance later associated with Peto and Haybittle?*[15]

PA: The response by Anscombe (who was, and continued to be, a good friend) reflected the growing Bayesian viewpoint. I was flattered that it publicized the book so well. I doubt whether the Bayesian view took firm hold with practical trialists, but I'm out of touch with current practice. As regards '3 or 4 standard errors': this implication of Anscombe's view was that one should sometimes continue recruiting even

though a very large difference, of say 3 standard errors, had emerged. My point was that if an interim analysis in a clinical trial had produced a difference of 3 standard errors, ethical issues might then be paramount. Peto and Haybittle were saying that stopping before that stage would be premature in not allowing for the effect of repeated sampling.

IC: *What happened to sequential analyses over the subsequent decade?*

PA: There was a trickle of sequential trials in the 1960s and a general awareness of the problems of repeated looks at data. In the second edition of *Sequential Medical Trials*[12] I introduced a few modifications and extensions, particularly in the use of 'repeated significance test' plans, where the stopping boundaries corresponded to conventionally significant results but at a higher significance level. However, my own involvement after 1976 became much more sketchy. Important later influences were the books by Pocock[16] and Whitehead.[17]

IC: *In 1978, you and Stuart Pocock analysed and reported a Union Internationale Contre le Cancer (UICC) survey of cancer trialists which revealed quite striking variations in practice.*[18] *What struck you most about your findings? Was this where the concept of 'group sequential designs' began to emerge?*

PA: The UICC had for some time had a working party on clinical trials, chaired by Daniel Schwartz, on which I served for many years. I am not sure whether this survey was initiated at the request of the working party, but I was not myself involved in its conduct, which was overseen by Stuart Pocock. The report shows that most of the trials used some form of statistical power calculation to determine trial size, and most used some form of interim analyses although only rarely with formal stopping rules. I don't remember my reactions at the time, but I suppose I would have been moderately pleased at the general outcome but a little disappointed that formal methods had not taken hold more firmly.

The idea of group sequential plans had emerged earlier. My own work had been based on the assumption that the results were analysed continuously, after each new patient's outcome was known. This was possible and acceptable with the typical small-scale trials reported in the 1950s and 1960s, usually under the control of one investigator. It was less appropriate for larger multicentre trials with analysts reporting periodically to data monitoring committees (DMCs). So the original plans were only a rough guide for use with group analysis. Pocock did a good job in presenting a theory of group sequential designs,[16] and this became widely used.

IC: *In 1979, in an article published in the* Australian Journal of Statistics *entitled 'The design of clinical trials', you discuss 'size of trials', noting:*

 (i) *'There are too many small trials, and "large" trials are not large enough.' (p. 272);*
 (ii) *'The determination of trial size at least partly from power considerations is clearly arbitrary and in no sense optimal.' (p. 272);*
(iii) *'I know of no trial that has been planned along decision–theoretic lines.' (p. 273);*

and on pages 273–274 you discuss sequential methods.[19] *Can you try to summarise where your thinking had reached at that point in time?*

PA: I suppose this was an attempt to summarise my thoughts about clinical trials a year or two after moving to Oxford and away from the front line of medical statistics. It comments on one or two general approaches that we haven't mentioned yet. Two of these concerned planned departures from randomization, either to balance risk factors ('minimisation') or to put more patients on the apparently better treatment ('play-the-winner', etc.). I was dubious about these, as they risked losing the benefits of randomisation. The 'play-the-winner' design achieved its purpose of putting more of the patients on the apparently better treatment but resulted in inefficient estimation of treatment differences because of the smaller number of patients receiving the apparently less effective one.

Another topic was the decision-theoretic approach to trial size, reflected in Anscombe's critique. Ted Colton,[20] in an elegant paper, had examined a 'horizon' model for clinical trials. The model postulated that one of two treatments was to be applied to a known population of patients (the 'horizon') and that the choice between the two treatments should be determined by a randomised trial (RCT) on an initial subset. The question is: how many subjects should be in the initial trial, leaving the rest to be given the apparently better treatment? Clearly, not too few, otherwise the wrong treatment might easily be chosen; but also not too many, because that would mean too many patients in the RCT having the worse treatment. Colton found that the trial should not involve more than one-third of the population at risk. Again, I was, and remain, dubious about the value of such models in the real world (as, I gather, is Ted Colton), particularly after my experience serving on data monitoring committees.

IC: *By 1980, you entitled a section of your paper in* Thrombosis and Haemostasis *'The Development of Large Trials'.*[21] *Was it theory/logic or examples that led you to emphasise the need for much larger trials?*

PA: It seemed an appropriate point to make for an audience of cardiologists – cardiovascular disease trials may involve follow-up with low event rates and perhaps small treatment effects that are nevertheless worth having. I had of course been impressed by Peto's advocacy of large simple trials in cancer. So, theory, logic AND examples were important!

IC: *In 1983, you gave a talk for the Society for Clinical Trials (published the following year in* Controlled Clinical Trials*), and reported that 'the case for some form of sequential analysis of data from clinical trials is widely accepted on ethical grounds, and the trend has moved away from fully sequential designs to group sequential designs for interim analyses, particularly for multicentre follow-up studies in which the appropriate committees meet at regular intervals'.*[22] *Had you become convinced that 'fully sequential trials' were no longer needed? When did data monitoring committees become usual?*

PA: We've covered some of this ground earlier. I don't think I would have ruled out fully sequential designs if the circumstances seemed to permit this, for example for a small trial being done in a single centre. But most of the trials I now heard about were multicentre, with interim analyses and data monitoring committees.

Data monitoring committees became common in the 1970s, especially for multicentre trials. Meinert's 1986 book has a list of many in the USA.[23] The earliest may have been that for the contentious Coronary Drug Project, which was set up in 1968. I served on an early DMC for a trial of heart disease prevention which started in 1971,[24] and on many others during the 1980s and 1990s.

IC: *It's in that 1984 paper that I think you first discuss 'Combination of trial results', viz. 'It becomes increasingly important to draw other conclusions from the whole body of data rather from each individual trial in isolation'.*[22] *You refer to John Lewis' analysis of the beta-blocker trials in particular.*[25] *Can you describe the development of your ideas about taking account of external evidence and meta-analysis in assessing when trials should stop recruiting?*

PA: I had no personal experience of meta-analysis/ systematic reviews, but became aware of the issue during the 1970s and 1980s through the writings of Tom Chalmers and Richard Peto. It seemed clear that reliable evidence from other studies should affect decisions about stopping.

IC: *Your 1989 paper 'Inference and decision in clinical trials' published in the* Journal of Clinical Epidemiology *has a substantial section on 'Planning the size and duration of a trial', referring particularly to Freedman and Spiegelhalter,*[26] *followed by a section*

of 'Early stopping'. Please summarise the main messages conveyed by those two sections.[27]

PA: In planning the size and duration of a trial it may be useful to define an 'indifference zone' around zero for a treatment difference, so that the trial need not be stopped merely because a zero effect can be contradicted.[28] This leads to questions as to what the limits of the indifference zone should be, and investigators may have different views about this. Freedman and Spiegelhalter and others have discussed various ways in which the prior views of the investigators can be elicited. However, they may not agree, and this may lead to different stopping decisions for different centres.

Various people have written about methods of 'stochastic curtailment' by which it may be possible during the course of a trial to predict fairly safely the final result, as either a likely definite difference or with the probability that no substantive difference will be detected (a situation sometimes called 'futility', with the suggestion that the trial may as well stop recruiting at this intermediate stage).

IC: *Your 1991 paper in* Statistics in Medicine – *'Interim analyses in clinical trials' – seems a key summary of the views you had reached by the early 1990s – group sequential analyses assessed by independent data monitoring committees.*[29] *Is that a fair summary?*

PA: Yes, but I didn't think that the analyses were the only relevant factors in deciding whether or when to stop. We'll come on to that in a minute. I was also starting to think that no model specifying the timing of the repeated analyses was ever likely to be exactly right and that the theoretical results should be regarded as general guidance rather than mandatory instructions.

IC: *At one point in the paper*[29] *you refer to the need for DMCs to take account of external evidence. Can you describe how your views on this were evolving?*

PA: For most or all the DMCs I served on in the 1980s I made no attempt to apply a strict sequential plan, largely because the pattern of interim analyses was difficult to predict. I tended to feel that this was likely to be a general situation, and I became less interested in formal rules. I was also aware that a termination decision could depend on other evidence – from other trials or research findings, adverse effects, administrative problems, etc. – which could not be predicted and put into the model for interim analyses.

My main experience of the use of external evidence was in the DMC of Concorde, the Anglo-French trial of zidovudine for HIV infection.[30] A concurrent American trial with a similar protocol, ACTG019, had been terminated early because of a reduction in early progressions in the actively treated group. The US investigators informed us of this at an early stage and visited the UK after their trial finished. The Concorde DMC recommended that our trial should not terminate yet, although our results went in the same direction, on the grounds that with longer patient follow-up the effect might disappear. This was seen later to have been a wise decision because our concerns proved justified by later analyses. It was a good example of the danger of inferring too much from short follow-up periods.

IC: *Your 1992 paper in the* Canadian Journal of Statistics *discusses frequentist and Bayesian approaches to stopping rules.*[31] *Am I right to be surprised that neither camp has emphasised the need to take account of plausible treatment differences as derived from meta-analyses of evidence from other trials?*

PA: I think you're right. I suppose either camp would consider amalgamating the evidence derived from the external and internal data at an interim stage (although not necessarily amalgamating the actual data), with some reservations about the relevance of the external work to the current study. Bayesians, in formalising degrees of belief, might have difficulty in quantifying this relevance.

IC: *At the end of the 1990s, you reported on the work of the Data Monitoring Committees for three major trials in HIV/AIDS – Concorde, Alpha, and Delta.*[30,32] *Can you summarise the main lessons you draw from that experience?*

PA: That's difficult! I have mentioned earlier the very interesting contacts with the US group during the monitoring of Concorde. The whole experience was fascinating and instructive, involving French colleagues from INSERM and elsewhere and input from specialists from different medical fields. We were served by a superb MRC data-processing team. My main impression was how thoroughly everything was reported, analysed and discussed. This experience gave me the opportunity to try out and modify general ideas about the statistical presentations, and, after the trials had been concluded, to write papers about the data monitoring procedures.[30,32] Although a difference of 3 standard errors was used as a guide as to when recruitment might cease, we adopted a very pragmatic approach, with few formal rules.

IC: *In conclusion, please would you try to summarise the evolution of your views about ways of deciding when clinical trials should stop recruiting.*

PA: That's even more difficult! My active involvement in methodology stopped in the 1970s, and contact with research groups became weaker, although I enjoyed the many DMCs I worked with. Since

around 2000 I have had little continuous involvement and have not attempted to keep up with literature. So my understanding of current practice and attitudes is very sketchy. My general impression is that the fusion of methodology and practical experience had, by 2000, led to sensible and acceptable procedures, relying more on common sense and spread of information than on technical rules. If I had to choose now, I would say that trialists should put primary emphasis on good practice – secure random assignment and unbiased assessment, careful observation and recording, etc. – rather than technical statistical procedures. Without the former the latter are meaningless.

IC: *Thank you, Peter, for sharing your reflections on half a century of thinking about how to decide when clinical trials should stop recruiting*.

Declarations

References

1. Wald A. *Sequential Analysis*. New York: Wiley, 1947.
2. Armitage P. Sequential analysis with more than two alternative hypotheses and its relationship to discriminant function analysis. *J R Stat Soc B* 1950; 12: 137–44.
3. Armitage P. Sequential tests in prophylactic and therapeutic trials. *Q J Med New Ser* 1954; 23: 255–74.
4. Newton DRL and Tanner JM. N-acetyl-para-aminophenol as an analgesic: a controlled clinical trial using the method of sequential analysis. *BMJ* 1956; 2: 1096–9.
5. Bross I. Sequential medical plans. *Biometrics* 1952; 8: 188–205.
6. Snell ES and Armitage P. Clinical comparison of diamorphine and pholcodine as cough suppressants by a new method of sequential analysis. *Lancet* 1957; 1: 860–2.
7. Watkinson G. Treatment of ulcerative colitis with topical hydrocortisone hemisuccinate sodium: a controlled trial employing restricted sequential analysis. *BMJ* 1958; 2: 1077–82.
8. Robertson JD and Armitage P. Report of a clinical trial to compare two hypotensive agents. *Anaesthesia* 1959; 14: 53–64.
9. Armitage P. Restricted sequential procedures. *Biometrika* 1957; 44: 9–26.
10. Brown A, Mohamed SD, Montgomery RD, Armitage P and Laurence DR. Value of a large dose of antitoxin in clinical tetanus. *Lancet* 1960; 2: 227–30.
11. Armitage P. *Sequential Medical Trials*, 1st edn. Oxford: Blackwell, 1960.
12. Armitage P. *Sequential Medical Trials*, 2nd edn. Oxford: Blackwell, 1975.
13. Anscombe F. Sequential medical trials. *J Am Stat Assoc* 1963; 58: 365–83.
14. Armitage P. Sequential medical trials: some comments on F.J. Anscombe's paper. *J Am Stat Assoc* 1963; 58: 384–7.
15. Haybittle JL. Repeated assessment of results in clinical trials of cancer treatment. *Br J Radiol* 1971; 44: 793–7.
16. Pocock SJ. *Clinical Trials: A Practical Approach*. Chichester: John Wiley, 1983.
17. Whitehead J. *The Design and Analysis of Sequential Clinicial Trials*. Chichester: Wiley, 1983.
18. Pocock SJ, Armitage P and Galton DAG. The size of cancer clinical trials: an international survey. *UICC Tech Rep Ser* 1978; 36: 5–32.
19. Armitage P. The design of clinical trials. *Aust J Stat* 1979; 21: 266–81.
20. Colton T. A model for selecting one of two medical treatments. *Bull Int Stat Inst* 1962; 39: 185–200. (also found in *J Am Stat Assoc* 1963; 58: 388–400).
21. Armitage P. Clinical trials in the secondary prevention of myocardial infarction and stroke. *Thromb Haemost* 1980; 43: 90–9.
22. Armitage P. Controversies and achievements in clinical trials. *Control Clin Trials* 1984; 5: 67–72.
23. Meinert CL. *Clinical Trials. Design, Conduct, and Analysis*. New York: Oxford University Press, 1986.
24. Elwood P. The first randomized trial of aspirin for heart attack and the advent of systematic overviews of trials. *JLL Bulletin: Commentaries on the History of Treatment Evaluation*. 2004. See http://www.jameslindlibrary.org (last checked 12 November 2013).
25. Lewis JA. Beta-blockade after myocardial infarction – a statistical view. *Br J Clin Pharmacol* 1982; 14: 155–215.
26. Freedman LS and Spiegelhalter DJ. The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *Statistician* 1983; 32: 153–60.
27. Armitage P. Inference and decision in clinical trials. *J Clin Epidemiol* 1989; 42: 293–9.
28. Meier P. Statistics and medical experimentation. *Biometrics* 1975; 31: 511–29.
29. Armitage P. Interim analysis in clinical trials. *Stat Med* 1991; 10: 925–37.
30. Armitage P. Data and safety monitoring in the Concorde and Alpha Trials. *Control Clin Trials* 1999; 20: 207–28.
31. Armitage P. Some topics of current interest in clinical trials. *Can J Stat* 1992; 20: 1–8.
32. Armitage P. Data and safety monitoring in the Delta Trial. *Control Clin Trials* 1999; 20: 229–41.