Biomedicine Special Issue, 1978, *28*, 24-36.

CLINICAL TRIAL METHODOLOGY

by Richard Peto

(Reader in Cancer Studies, Radcliffe Infirmary, Oxford University*, Oxford)

INTRODUCTION

The essential attributes of a good clinical trial are that it should address itself to an important question, get the correct answer, and be convincing to other workers when published. In my opinion, once an important question such as this is being considered, the best way to achieve these ends in answering it is usually to undertake a comparative study which is :

i. *Large*, for in most current trials random differences between groups could well swamp medically important effects or could masquerade as important effects. Moreover, although the selection of large trials presented in the world literature is a moderately biased sample of all large trials (« accentuate the positive, eliminate the negative, latch on to the affirmative », etc.), these biases are slight in comparison with the selective biases that determine which *small* trials are published or, still more, that determine which small trials are presented at meetings. Experience of these biases has led many people to distrust the results of small trials, and thus large size is necessary not only to be correct but also to be convincing. Appendix 1 discusses reasons and methods for getting large trials.

ii. *Randomised*, for this should ensure that only random differences can affect the treatment comparison. By contrast, when comparing one series of patients with a previous or other series of supposedly similar patients, some systematic non-random biases could well exist, and in many cases these biases are large enough to be medically misleading.

(Stratified allocation is unnecessary but harmless, except that it is a slight nuisance. Factorial randomised

* Oxford OX2 6HE, England.

designs such as 2 × 2 designs may be much better than ordinary 2-way randomisations; see Appendix 2 for definitions and details of 2 × 2 designs.)

iii. *Terminated properly* by criteria previously outlined in the protocol – although it is, of course, not always practicable to formalise in advance all of the subtle determinants of when a trial should stop.

iv. *Reported without any exclusions*, for if « protocol deviants » or « unevaluable patients » are entirely omitted from an analysis, then bias may result. Exclusion of protocol deviants is perhaps the commonest serious error in large randomised trials.

v. *Analysed fully and correctly*, making extensive use of explanatory information recorded at entry on each patient. Far more than just one or two P-values is needed to understand the results of a clinical trial properly. (For example, in survival analyses I usually use logrank P-values for treatment effects and for the correlations of explanatory information with survival, illustrating these analyses with life-tables. Then I look at treatment effects in lots of particular subgroups, and also at treatment effects after retrospective stratification for certain explanatory information).

vi. *Interpreted cautiously*, for even if there is a medically important difference between two treatments, a non-significant result may well emerge from a trial comparing them, especially in a small trial. Conversely, in practice « $P < 0.05$ » treatment differences can unfortunately emerge by chance quite easily from trials comparing two equivalent treatments. This is partly because, while the chance of finding $P < 0.05$ *at any one particular time* should be 0.05, the chance of finding $P < 0.05$ *at some time or other* during the course of a trial is more like 0.2 (6). Also, if the apparent treatment effect in a particular subgroup (e.g. younger females) is much more extreme than the apparent treatment effects elsewhere, then although this fact should be noted it is likely to be much weaker evidence for real heterogeneity of response than it appears to be.

Let me explain why I feel that the real interests of, for example, cancer patients will be served best by the sort of studies I have just described.

For decades, resectable cancers have been removed surgically and cancer registry data and national mortality data indicate that during these recent decades survival from the major cancers has in general not materially improved, despite the concurrent development of medical oncology. These are historically controlled comparisons on a grand scale, and like most such comparisons they are open to various interpretations. My interpretation is that, having done what resections we can, almost all past claims or hopes of great therapeutic improvements have been mistaken, and so, despite appearances, almost all current therapeutic suggestions will likewise eventually be found to yield either small, or no, benefits. (Small but definite improvements in the treatment of common cancers may be very important, of course.) *Rigorous and sceptical trials designed to distinguish between small benefits and no benefits are therefore needed :* if, exceptionally, the benefits of a particular treatment are remarkably large, then what was intended to be a large trial can always be aborted once *really* definite evidence has accumulated. Such trials are needed (*a*) to test new claims, and (*b*) to investigate which current treatments are literally useless, especially the more toxic or expensive current treatments. (Trials of expensive treatments versus much cheaper treatments are also desirable.) Because the need is to distinguish between small benefits and no benefits, *historically controlled* comparisons will not suffice, nor will *small randomised* trials suffice; for in both cases the possible errors inherent in the methodology exceed the likely magnitude of the differences being characterised (Treatment differences in clinical trials will usually be small not only in diseases where therapeutic progress continues to be slow but also, perhaps surprisingly, in diseases such as acute leukaemia or Hodgkin's disease where very substantial progress has been made or is being made. This is because the « great progress » in these diseases typically consists of a few major innovations plus many improvements in supportive care, and most trials, even in these diseases, are not comparing a great improvement with its predecessor; instead, most trials have been comparing rather similar treatments, either both new and good or both old and bad.)

The careful studies that are now needed will be impeded unless more trial organisers and their statisticians become aware of the usual need for really large intake to a trial; the practical difficulties of attaining the numbers that one would ideally like are formidable, even for the commoner tumours, but considerable improvement on the trial sizes that are currently commonplace would not be difficult, and Appendix 1 consists of sixteen particular suggestions of devices for getting larger trials. In my opinion, greater awareness of the need for large numbers, and development of collaborative organisations which will bring together large numbers of patients, is the single most important improvement in trial methodology now possible. I view the needs for randomisation, proper termination, full inclusion of every randomised patient, thorough analysis, and cautious interpretation as being obvious and straightforward by comparison;

in general, one usually only has to explain to trial organisers or statisticians the possible biases if these needs are ignored for them to be agreed and implemented.

ALTERNATIVE METHODOLOGIES : HOW LARGE ARE REAL TREATMENT EFFECTS LIKELY TO BE?

The difficulties in setting up randomised studies of adequate size are so formidable that, not unnaturally, people have tried to devise viable alternative ways of evaluating treatments. However, one must be very cautious, for although the history of medicine contains many examples that demonstrate that some, and probably nearly all, *substantial* innovations can be recognised without the support of randomised trials, it also contains many examples of expensive, unpleasant or harmful practices that have been mistaken for substantial innovations — super-radical mastectomy, harmful immobility after myocardial infarction, miserable bland diets for patients with ulcers, portacaval shunts, beta blockers to delay labour, some coronary artery bypass surgery, most physiotherapy, most bronchodilators in emphysema, and much of the treatment of lunatics. (Also, we do not know how many moderate improvements have been unrecognised and laid aside because of inadequate testing, but the history of lithium treatment for manic depression provides one instance of this which was not rectified for 20 years.)

Two great absorbers of medical finance nowadays are chemotherapy for carcinomas (which probably does more harm than good for many categories of patient), and admission to coronary care units following a myocardial infarction. Coronary care units offer an interesting parallel which oncologists can perhaps consider unemotionally. Since their inception, dozens of improvements in the speed with which patients get rushed into them, the monitoring techniques, the resuscitation techniques, and so on, have been suggested, discussed, developed, and adopted. Heart disease is the major cause of death in developed countries, and almost any cardiologist can tell stories of patients who were brought back from death in his unit and who later recovered fully. However (despite accusations of unethicality), Mather *et al.* (5) decided to set up a randomised trial to see whether coronary care units really did save lives. 450 patients with a recent infarct were randomised equally between admission to coronary care units and remaining at home. In the event, 71 deaths occurred in hospital and 64 at home. This does not show that coronary care units are completely without benefit, of course, because the trial only had 450 highly selected patients in it and so was not large enough. (In a less selected but smaller such trial, Hill et al. (4) have observed 14 deaths in hospital and 17 at home.) Indeed, categories of patient almost certainly do exist for whom admission to coronary care is advisable. This trial does, however, illustrate that treatments which are thought on good *a priori* grounds, but in the absence of proper evaluation, to be therapeutically important *may* be of little or no true value to particular, identifiable categories of patient to whom they are currently given.

The relevance of this observation to clinical trial methodology is that it re-emphasises the point, which I have already made in the context of oncology, that *useful trials must usually be capable of distinguishing between the alternative possibilities of a small treatment effect and no treatment effect.* It is obviously only possible to do this with confidence if both the *random* and the *systematic* errors in one's trial methodology are smaller than the difference between no effect and a small but interesting effect. Randomised trials with objective stopping rules have no systematic errors, but, unless they are large, they suffer from serious random errors. If historically controlled studies, because they do not have to set aside any patients as controls, contain a larger group of treated patients, this could reduce these random errors, but this is at the expense of introducing systematic errors of unknown magnitudes in the selection of controls, the selection of patients, the definition of the beginning and end of the control period, the exclusion of inadequately treated patients, the stopping rule, and so on. Even the *organiser* of a non-randomised trial can usually only have a hazy idea of the possible magnitudes of his systematic biases, and the sceptical reader must be even more uncertain of them, so even if it is correct a historically controlled study may be unconvincing. For these reasons, I think that real medical progress will be more rapid if ponderous, slow, difficult, large randomised studies are encouraged at the expense of small, rapid, practical, historically controlled comparisons.

Any judgement such as this which suggests how overall medical progress might be accelerated will be judged not only by the theoretical arguments advanced in its support, but also by considering the history of medical progress. However, this is more difficult than it seems, not only because one naturally examines the ways in which real progress has been made (which are by definition the times when the existing system worked adequately), but also because we have only got *one* history to examine. We can't run history through twice, first starting in 1930 with one philosophy and then restarting the world in 1930 with another philosophy, to see where the alternative

world-lines would have led us. Thus, although twentieth-century medical research has achieved many things, perhaps with better controlled treatment evaluations we could have achieved even more. An instructive example which suggests that this may indeed be the case is provided by the history of anticoagulant therapy for myocardial infarction (MI).

« Phase III » studies of anticoagulant therapy for MI

Chalmers *et al.* (3) have recently reviewed all 32 comparative studies of anticoagulant therapy, involving a total of 16000 MI patients, published over the last 30 years and much of the data they compiled appears in Table I. Superficially, these trials suggest that historical controls suffice : of 24 non-randomised comparative studies in which the mortality results in the two groups differed, all but two non-significant exceptions differed in a direction favouring anti-coagulants, and many of these differences were highly statistically significant. (The largest such study reported 27 per cent mortality in 1387 historical controls as against only 8 per cent mortality in 841 anticoagulated patients, corresponding to a P-value of less than one in a million trillion.) Three large randomised trials were eventually (1969, 1972, 1973) reported, and in each randomised trial somewhat fewer of the anticoagulated patients died. When the randomised trials were pooled, a statistically significant ($P < 0.01$) difference was observed, thus confirming the conclusions previously suggested by the historically controlled studies that anticoagulation is wise. However, the obvious troubles with the simple interpretation that « historical controls gave the right answer » are :

1) The non-randomised comparisons were not convincing to many physicians during the 1950 s and 1960 s, and so for 25 years a practicable treatment which could have saved the lives of a proportion of MI patients was not generally adopted. Even now, perhaps because of past confusion, some physicians still feel doubts as to whether routine anticoagulation is of value. (If these doubts persist, however, then the ability of randomised trials to impress physicians will also be uncertain, here as in diabetic therapy.)

2) If all the *randomised* studies are pooled, the average magnitude of the apparent protective effect is 20 per cent (i.e. 20 per cent of deaths are prevented, $\chi^2_1 = 7.4$, and a 95 % confidence interval for the proportion prevented is approximately 5 %-35 %, so that we can be fairly sure that between one twentieth and one third of all deaths can really be prevented.) If all the *non-randomised* studies are pooled, the average magnitude of the apparent protective effect is 53 % (95 % confidence interval 46 %-60 %) so that apparently just over half the deaths were prevented. This discrepancy shows that the average bias in the historically controlled studies was of the same order of magnitude as the effect being studied, and in many cases (e.g. the largest study, where 27 % of historical controls died against only 8 % of treated patients), the biases must have substantially exceeded the real effect being studied. In other words, these historically controlled studies would have yielded positive answers on average even if anticoagulation were slightly harmful and wholly without benefit. Such methodology is of little value when, as here, the benefits are only moderate (but nevertheless important).

3) To be fairly sure of detecting a 20 % improvement in mortality, a randomised trial involving about 3000 MI patients would be necessary, and lesser effects would require even larger numbers. (Strictly, the sensitivity of a study depends on the number of deaths, not the total number of patients. The 500 deaths that one might see in 3000 such patients could perhaps be observed in only 1000 patients suffering from a more dangerous disease than this.)

4) Due to their small size, many of the studies, randomised and other, would have been very unlikely to yield a significant effect if the true improvement really is of order 20 %, and moreover even if they had (e.g. if the trial with 31/92 dead had happened to yield 21/46 and 10/46, which just gives $P < 0.05$), the magnitude of the apparent difference would then have been misleadingly great.

5) Three of the 6 randomised studies were so small as to be of almost no scientific value.

6) If there was any bias in the concurrent non-randomised studies, then one might expect the proportion of deaths prevented to be greater in them than in the randomised studies. This was indeed the case, although the difference (27 % vs. 20 %) is not statistically significant.

7) At a distance of some years, the scientific evidence from the randomized trials remains solid and usable, while doubts which either could never be answered or which cannot now, at this distance in time, be answered accumulate around the other studies.

These anticoagulant trials show that the errors inherent in the sort of historical control series which are chosen in practice can easily be of the same order of magnitude as the effects which one might commonly wish to study in a clinical trial. Of course, that does not prove that historical controls have to be as bad as this. Perhaps more care in the selection of historical controls could have considerably reduced

TABLE I

Results of 32 comparative studies of anticoagulation therapy for MI patients

Reference in Chalmers et al (3)	Year of publi-cation	Total size of study	No. of control deaths	No. of control patients	Per cent controls dying	Observed No. of treated deaths	No. of treated patients	Per cent treated deaths	Expected No. of treated deaths	Variance of (observed-expected)	Ratio of percen-tages
18 Studies employing historical controls (pooled Mantel-Haenszel chi-square from 18 studies for treatment effect = 302.0 on 1 DF)											
(2)	1948	175	35	100	35.0	7	75	9.3	18.0	7.9	0.27
(3)	1951	920	187	731	25.6	27	189	14.3	44.0	26.8	0.56
(4)	1953	311	68	211	32.2	18	100	18.0	27.7	13.6	0.56
(5)	1953	200	51	125	40.8	19	75	25.3	26.3	10.7	0.62
(6)	1953	31	4	12	33.3	7	19	36.8	6.7	1.7	1.11
(7)	1954	745	75	242	31.0	69	503	13.7	97.2	25.5	0.44
(8)	1956	314	45	150	30.0	23	164	14.0	35.5	13.3	0.47
(9)	1957	362	38	90	42.2	63	272	23.2	75.9	13.6	0.55
(10)	1957	543	136	284	47.9	59	259	22.8	93.0	31.2	0.48
(11)	1958	264	93	198	47.0	31	66	47.0	31.0	12.4	1.00
(12)	(1958)	(436)	(44)	(110)	(40.0)	(67)	(326)	(20.6)	(83.0)	(15.6)	(0.51)
(13)	1961	308	16	68	23.5	41	240	17.1	44.4	8.0	0.73
(14)	1962	128	8	28	28.6	11	100	11.0	14.8	2.8	0.39
(15)	1962	191	37	68	54.4	34	123	27.6	45.7	10.3	0.51
(16)	1962	104	8	17	47.1	25	87	28.7	27.6	3.1	0.61
(17)	1964	674	24	76	31.6	87	598	14.5	98.5	9.3	0.46
(18)	1975	2 228	379	1 387	27.3	67	841	8.2	168.4	83.9	0.29
(19)	1975	1 156	177	673	26.3	52	483	10.8	95.7	44.7	0.41
Sub-totals excluding (12)		8 654	1 381	4 460	31.0	640	4 194	15.3	950.4	318.9	0.49
(Weighed averages of percentages dead)					(31.6)*			(14.9)*			(ratio = 0.47)
8 Studies employing concurrent non-randomised controls (pooled Mantel-Haenszel chi-square from 8 studies = 25.2 on 1 DF)											
(20)	1948	800	88	368	23.9	65	432	15.0	82.6	30.8	0.63
(21)	1950	250	16	128	12.5	23	122	18.9	19.0	8.3	1.51
(22)	1950	154	34	84	40.5	16	70	22.9	22.7	8.4	0.56
(23)	1951	430	92	256	35.9	39	174	22.4	53.0	22.0	0.62
(24)	1952	152	23	76	30.3	23	76	30.3	23.0	8.1	1.00
(25)	1952	332	57	171	33.3	36	161	22.4	45.1	16.8	0.67
(26)	1959	226	45	115	39.1	21	111	18.9	32.4	11.7	0.48
(27)	1961	800	109	429	25.4	85	371	22.9	90.0	36.6	0.90
Sub-totals		3 144	464	1 627	28.5	308	1 517	20.3	367.9	142.6	0.71
(Weighted averages of percentages dead)					(28.2)*			(20.5)*			(ratio = 0.73)
3 Small randomised controlled trials, none statistically significant											
(28)	1960	92	18	47	38.3	13	45	28.9	15.2	5.2	0.75
(29)	1966	147	15	70	21.4	12	77	15.6	14.1	5.5	0.73
(32)	1972	53	2	26	7.7	2	27	7.4	2.0	0.9	0.96
3 Large randomised controlled trials (pooled Mantel-Haenszel chi-square for all 7 lines of data from randomised trials = 7.4 : P < 0.01)											
(30)	1969	1 427	129	715	18.0	115	712	16.2	121.7	50.6	0.90
(31)	1972 (*a*)	632	37	155	23.9	64	477	13.6	76.2	15.7	0.56
	(*b*)	504	46	236	19.5	47	268	17.5	49.5	18.9	0.90
(33)	1973	999	56	499	11.2	48	500	9.6	52.1	23.3	0.86
Total randomised		3 854	313	1 748	17.9	301	2 106	14.3	330.8	120.2	0.80
(Weighted averages of percentages dead)					(17.9)*			(14.3)*			(ratio = 0.80)

Note : the trial reported in ref. (31) consisted of (*a*) 3 years randomised 3 : 1, then (*b*) 3 years randomised 1 : 1.

* These percentages in brackets represent averages of the separate percentages, weighted in proportion to total trial sizes. Although typical, ref. (12) has been omitted from all totals because its control group was part of that for ref. (10).

these biases. It has been suggested, for example, that, in centres undertaking a series of trials with common entry criteria, groups of patients in successive trials could be fairly reliably compared with each other. However, my experience when a common treatment arm carried over from the fifth MRC acute myeloid leukaemia trial to the sixth (next) trial is that a surprisingly large ($P < 0.01$) difference arose, for reasons which are still unclear.

Likewise, in North America, Pocock (10) has collected 19 unselected instances where collaborative groups used the same entry criteria for two successive trials, and carried one particular treatment arm over from the first trial to the second, yielding two groups

of patients who should, if historical comparisons are reliable, fare rather similarly. However, in some cases these pairs of groups fared very differently; indeed, of the 19 significance levels comparing survival within such pairs, four were $P < 0.02$*.

If the second trial treatment had differed in some unimportant detail from the first and a « historically controlled » comparison had been made to see whether this detail mattered, these $P < 0.02$ results could have been very misleading. Historically controlled « Phase III » comparisons are clearly less reliable than might intuitively be expected, a conclusion which is reinforced by the excellent papers of Byar et al. (1) and of Chalmers et al. (2).

The possible role of randomisation in « Phase II » studies

The development of a new therapy (using surgery, drugs, radiation or special patient support) can sometimes be approximately divided into three « phases ». In « Phase I » the practical problems (e.g. toxicity, excretion, serum assay, maximum tolerated dose, etc.) associated with the new treatment are more or less sorted out, along with any organisational problems in using it. During Phase I, there is little or no deliberate assessment of the efficacy of the treatment; the aim is chiefly to devise practicable schedules, which is an inevitable first step.

In « Phase II », the treatment is then given to a number of patients (perhaps several hundred in all) to see which categories, if any, show some sort of slight response to the treatment (e.g. a temporary tumour shrinkage.) At present almost all such Phase II studies are non-randomised. This lack of randomised controls introduces the danger, if « response » is defined too loosely, of mistaking chance variations in the disease or in its assessment for real effects. It also introduces the even greater danger, if « response » is defined too rigorously, of failing to recognise a moderate therapeutic effect. (A new treatment is unlikely to be given optimally in a Phase II study, and so a really useful treatment may only have a moderate effect in Phase II.) Having identified, from the preliminary Phase II studies, a few categories of patient who may possibly show some response, later Phase II studies may chiefly concentrate on patients in these categories, hoping to estimate the percentage of them that are likely to respond to this new treatment. Again, most such late Phase II studies are non-randomised, and again this sometimes produces very considerable difficulties of interpretation.

These difficulties are apparent when, as Moertel & Reitemeier (8) did, we compare the results from 20 different such « Phase II » studies of the same agent (5 FU by rapid injection) for the same disease (advanced carcinoma of the large bowel). Their summary results are reproduced in Table II, where it may be seen that the percentage of « objective responses » varied between 8 % and 85 %! It seems probable that more reliable information could have been gained, not only about « objective responses » but especially about more marginal responses, by randomised studies on several hundred of these thousand or more patients.

Next, the treatment having been devised in Phase I and been found to be, at least in the opinion of those who have used it, moderately efficacious in Phase II, it is finally supposed to be evaluated in comparison with alternative treatments in controlled « Phase III » trials. Unfortunately, the results from the earlier uncontrolled Phase II studies may cause considerable ethical difficulties in Phase III. What if there is no alternative treatment, or what if the new treatment is such (e.g. a sterile environment for leukaemia patients, or an anti-oestrogen agent for breast cancer patients) that it does not interfere with any standard treatment modalities? It may then happen that, by the time Phase III studies are planned, the investigators who undertook the Phase II studies are so convinced of the value of the treatment that they cannot ethically deny it to the Phase III controls. This would mean that the people with most experience of the new treatment (perhaps including the team who developed it), and who may therefore give it best, cannot ethically participate in its objective evaluation. This would be an advantage if the uncontrolled Phase II studies which convinced the original investigators are scientifically reliable and convincing to other workers, for the whole world would then immediately adopt the new, superior treatment. However, uncontrolled Phase II studies may be misleading (for example, whether or not hyperbaric oxygen is of any benefit, it is certainly not as useful a radiosensitiser as it was first believed to be, and also some forms of immunotherapy have not proved as effective as was hoped. Thus impressive Phase II studies, even though convincing to those who undertook them, may not be convincing to sceptics elsewhere, and so, whether or not the sceptics are correct, the truth may take years longer than necessary to emerge. (An exception,

* Pocock (pers. comm.) gives the 19 2-sided P-values as $P = 0.0001$ $P = 0.0016$, $P = 0.010$ $P = 0.019$, .07, .12, .13, .18, .19, .19, .31, .52, .57, .60, .62, .62, .78, .91 and .98. No adjustment for explanatory information was made in computing these P-values, and it is not known what effects such adjustment might have. Explanatory information is not available for the $P = 0.0001$ comparison, while in the $P = 0.0016$ comparison explanatory information has been used, but has not sufficed to explain the discrepancy.

TABLE II

Results of 20 uncontrolled studies of rapid 5-fluorouracil injection for advanced colorectal cancer reviewed by Moertel and Reitemeier

Reference in Table 10.1 of Moertel & Reitemeier (8)	Year of publication	No. of Patients Treated	No. of objective regressions observed	Per cent objective regressions
7	1962	13	11	85
8	1962	19	12	63
9	1962	47*	26	55
10	1961	17	8	47
11	1960	13	6	46
12	1962	12	5	42
13	1963	37*	15	41
14	1962	22	8	36
15	1961	37*	13	35
16	1960	12	4	33
17	1964	150*	47	31
18	1967	48*	13	27
19	1964	183*	42	23
20	1962	30*	6	20
21	1963	141*	24	17
their own series	1969	144*	22	15
22	1962	87*	10	11.49
23	1964	22	2	9
24	1961	11	1	9
25	1960	12	1	8
Total		1 057	276	26

Chi-square for heterogeneity = 107.54, DF = 19, $P < 0.0001$.

* These are the ten largest studies. The mean response rate in the ten smallest studies (58/153 = 38 %) was significantly higher than the mean response rate in the ten largest studies (218/904 = 24 %), perhaps because the small studies were selected differently or reported prematurely because the early results from them were rather extreme. (However, even among the ten largest studies the percentages of responders were grossly heterogeneous, ranging from 11 % to 55 %.)

of course, is the situation such as kidney transplantation, heart transplantation or bone marrow transplantation, where any long-term successes suffice to demonstrate that the proposed treatment can work and where randomised controls would be wholly irrelevant until, later on, we try to discover which categories of patients need such treatments.)

A practical compromise — 2 : 1 randomisation in Phase II

There are thus sound reasons for preferring randomised studies (even, perhaps, quite small randomised studies) in place of many present « Phase II » non-randomised studies. Indeed, for certain treatments one might start so early with some scheme of randomised control that there might be almost nothing left that would be called « Phase II ». This would be counter to the instincts of an investigator who has just devised a new treatment and who now wants to try it out on everyone he possibly can, and perhaps some compromise between this investigator's desire to treat everyone and his statistician's desire to leave half of them without the new treatment would be optimal.

If an investigator randomises two-thirds of his patients to his new treatment, leaving one third as randomised controls, then this combines most of the advantages of a randomised study with most of the advantages of a historically controlled study. Most of the patients will get the new treatment and can be compared with « experience », or a historical series, or whatever, but as a corrective (or a stimulus) to optimism there will be an unbiased randomised comparison which is very nearly as sensitive as an ordinary equal-groups randomised trial on these numbers would have been. Often there is no scientific sense in which a purely historically controlled Phase II study is superior to such a 2 : 1 randomisation, so why not adopt this compromise? I hope that it will be possible to proceed thus when interferon becomes available for British cancer trials in 1978-9.

ETHICAL QUESTIONS

There are ethical difficulties in clinical trials, especially as moderately strong evidence in favour of one treatment begins to emerge, and I would strongly support Chalmers' suggestion that the decision as to when a trial should stop should be in the hands of a small supervisory committee, containing both statisticians and physicians, none of whom are themselves entering patients into the study. Their considerations can be wider than those of the physician who is confronted with the patient and his family. Let me try to explain this euphemism more fully. The supervisory committee may well be prepared to let a particular trial go on until a treatment difference of well over two standard deviations emerges. This would depend on the available literature, on the likely response to the trial if published as it stands, and so on. If a trial is topped prematurely when it looks promising then other people will mount trials to try to settle the issue — remember the 32 trials of anticoagulants on 16,000 MI patients — and in the end it is likely that more patients will eventually get admitted to such studies than if the original study had been allowed to run on well beyond two standard deviations, perhaps to 2.5 standard deviations or more. This is to a rather limited extent, sacrificing present patients for the sake of future ones. I say « to a rather limited extent », because I suspect that in most trials the treatments will be of rather similar efficacy (and the most serious problem with ethics that is then likely to arise is how to avoid the whole project becoming bogged down in ethical minutiae). Moreover, the mere fact of participation in a trial may improve the standard of medical care enough for both treatment groups to fare better than they would have if similar treatments had been administered in a non-trial context.

I do not consider it ethical to impose such restrictions on the practice of clinical trials that the objective evaluation of treatments which do not differ substantially becomes impossible, for this would serve the real interests of almost nobody; indeed, I would go further and say that where there is doubt as to what treatment is best, there is an ethical need to act in such a way that it is discovered how patients can best be treated. There is, however, a very delicate balance between the need for knowledge and the needs of the patient, and neither can be put aside. Particularly,

1) If, when considering a patient for entry into a trial, the physician feels (either for medical reasons, or because of the patient's definite wishes, or for any other definite reasons) that at least one particular one of the possible treatments should not be given to that patient, then that patient should not be entered and should be treated however the physician thinks best.

2) If, during a trial, the physician feels *strongly* that a particular treatment is indicated or contraindicated, then the patient must be treated as thought fit, no matter what treatment arm of the trial he is nominally in. However, this only applies to strong preferences. (Since we don't want *too* many such protocol deviations, physicians with a definite propensity for strong preferences might be encouraged not to participate.)

If, by reasonably sensitive questioning before entry to a study, the physician decides that a particular patient whould object to one or other treatment, or has other relevant strong views, then these should be respected and that patient should be excluded. In cases of unusual doubt, it is worth soliciting informed consent prior to randomisation, but there is no good ethical reason whatever to solicit informed consent from all patients. Much of the edifice of informed consent is a legalistic trick to devolve what should properly be the doctor's responsibilities onto the patient. It may serve a useful purpose in warding off American lawyers, but in less litigious countries it is not necessary unless the doctor concerned feels it to be so. Indeed, there are occasions when soliciting informed consent is definitely unethical, as it may cause the patient unnecessary wory (e.g. when evaluating treatments for possible occult disease in patients who are clinically well). The Medical Research Council of Great Britain made this very clear in 1964 in their statement on Responsibility in Investigations on Human Subjects (7), a document which still stands as a guideline to medical research in Britain.

APPENDIX 1

Reasons and methods for getting large numbers of patients into a trial

Trial size : 1) Necessity.

The first essential in getting large numbers is for the trial organisers to be convinced of the urgent need for large numbers. If they are not already convinced of this, the statistician's first and major task must be to make them acutely aware of this need. How can one persuade an investigator that he really needs much bigger numbers than he thinks he does to compare the effects of two treatments on survival? (NB If the intended endpoint is not survival but instead some change in a biochemical, physical or other measurement related to the disease, then quite small numbers of randomised patients may indeed suffice.) Examples of results that might emerge by chance alone if there is *no* real difference between the two treatments do not seem to be as shocking to clinical investigators as examples of how really important differences could easily be missed in the trials they plan.

For example, if someone wants to randomise 100 Stage II breast cancer patients to chemotherapy or control, one might say « Look, suppose that your chemotherapy is really so great that it actually *cures one third* of the patients who would otherwise have died. That would be really important medically, but with only 50 patients in each group you could very easily miss this. For example, you could well finish up with 21 deaths of treated patients and 24 deaths of control patients (which essentially looks like a negative result), instead of the 18 : 27 you expect. »

Actually, even with 1000 patients it is possible for a medically important difference not to be statistically significant, although the actual results will then almost certainly at least point substantially in the right direction.

Example. Consider the comparison of two chemotherapeutic regimes for Stage II breast cancer, Intensive (I) versus Moderate (M). Suppose that the 5-year probability of death after I is 0.4 and after M is 0.5. (This difference matters.) Suppose that the trial stops and is reported when 360 patients have died. We would expect these 360 deaths to be distributed 160 I : 200 M ($P = 0.01$), but they could easily turn out by chance to be 170 I : 190 M instead, which would not be anywhere near statistical significance ($P = 0.2$).

Trial size : 2) How can big numbers actually be achieved?

First and most important, can collaboration between several centres be achieved, so that the trial

intake is not limited to the patients referred to one single clinic? Collaboration need not involve you organising collaboration of other clinics in the trial which you envisage; it could instead involve you collaborating with someone else's trial (probably in your own country), or you making your trial sufficiently similar to someone else's trial (in your own country or abroad) for the two to eventually be viewed together. These last two possibilities should be considered seriously far more often than they usually are.

Trial size : 3) Collaboration with other trials.

When you wish to organise trials in a particular disease, or trials of a particular treatment, do not do so in ignorance of the trials in that disease or the trials of that or similar treatments that are being organised elsewhere, especially in your own country. A couple of weeks delay while you correspond with physicians, chiefly in your own country, who are already engaged in planning or running such studies (or who are likely to know who is thus engaged), and with, in Britain, the Medical Research Council, or, in America, the appropriate branch of the NIH may yield useful perspectives. (The UICC and the NCI keep up-to-date lists of ongoing cancer trials.)

If related studies do exist or are being planned, it may well be that the interests of medical knowledge would be served better by cancellation of your plans to set up your own study and, instead, collaboration with another group. Physicians who could run their own small studies but who choose instead to collaborate with other studies are often worthy of considerable respect, for they are in many cases consciously, for the benefit of serious medical research, foregoing the immediate professional prestige associated with running a study themselves.

If strongly related randomised studies do exist but for good geographic, personal, or other reasons, full collaboration is undesirable, most of the benefits of full collaboration may yet be achievable if you organise a study which (although it may have moderately different entry criteria) compares treatments which are similar to the treatments being compared elsewhere. This is because an overall analysis of the two studies will then eventually be possible, using retrospective stratification. (For example, the several different randomised studies of post-operative radiotherapy for Stage II breast cancer could be pooled thus, yielding as accurate an estimate of the effects on mortality as would have been obtained in a single randomised study of several thousand women. Likewise, the six randomised studies of anticoagulant therapy for MI listed in Table I can be pooled, as is done in Table I.)

Trial size : 4) Establishing a trial with which other centres may collaborate

i. *Status of collaborators*

You must do all that is possible to motivate collaboration, for the extent of collaboration may well be the single most important determinant of whether the trial succeeds or fails. Establish a working party with wide representation, have regular meetings of potential or actual participants at which the main scientific issues (and relevant results from trials elsewhere) are presented and discussed as well as the problems of your own trial. Always emphasise that the trial is the property of all the participants, not the property of the trial centre, and that this will be reflected in all publications deriving from it.

ii. *Simple forms*

Documentation should be made as simple as possible : *really minimal* documentation at entry and during follow-up might increase the willingness of other physicians to participate. The statistician should, at the design stage, cross out from the draft coding forms most of the things that the trial organiser thinks he wants to ask! (Also the layout should be simple, with clearly written indications of what is required and all the places for answers in an obvious sequence down the right hand side, not all over the place.)

At entry we need the full name, sex and exact date of birth (for follow-up using national archives), we need the result and date of randomisation (to start survival analysis from this date), and we need to know those *few* prognostic features which correlate *strongly* with prognosis (to facilitate retrospective stratification, and to help define subgroups in which one treatment is clearly better and subgroups where it is not). Collection of more data than this may clarify the clinical course of each patient, but it is not a necessary part of a randomised trial and should usually be avoided.

Likewise, during follow-up, one need only ask whether and when the chosen endpoints have occurred (alive/date and cause of death; still recurrence-free/date of recurrence), what treatment side-effects have occurred, and whether the allocated treatment has been administered. Because one often does not know what the « side effects » might be, it may be advisable to ask that *all* medical events be noted. (In large trials, it might suffice to record whether or not the treatment has been administered only in a random subsample.) These simple follow-up data should be solicited regularly (and insistently!) but infrequently.

iii. *Accept anybody*

« Centres of excellence » often say of certain other centres « We don't want them; they're not good enough and would spoil the trial ». This is often nonsense; as long as analysis involves retrospective stratification (which it should, if only to avoid the need for stratified randomisation), patients at one centre need never get compared with patients elsewhere but only with each other. Also, it's easy to do a special analysis including only certain pre-specified centres that think themselves excellent, so nothing can be lost by widening intake.

iv. *Make complexity optional*

Because a trial involves the collection of a larger series of patients with a particular disease than may commonly be available, special unrelated studies (e.g. of particular biochemical effects or minor disease endpoints) are often planned for the trial patients. Although such studies may eventually prove to be of more scientific value than the trial itself, to be of value they may not need nearly as many patients as the trial needs. The statistician should therefore try to get the planned data items sorted into *trial* items and *other* items. If it is impraticable to recommend extension of the complete study to many centres, the statistician should consider recommending that just the essential simple randomised trial be extended without extending the ancillary studies. (Equivalently, if a trial is planned *ab initio* as a multicentre collaborative project with ancillary investigations, the statistician can recommend that the ancillary investigations be made completely optional, with no pressure whatever to do them.)

v. *Simplify treatment*

Treatment regimes should, if possible, be simple and easy to administer, rather than intricate and difficult. If one detail of treatment can be changed to make one treatment more widely practicable, that change should be considered. (This has the added advantage of making the results of the trial more widely relevant.)

vi. *Money*

Should all clinicians be offered by return of post £10 in cash (for departmental funds) per new patient randomised, as compensation for « secretarial expenses »?

vii. « *Flag* » *all the patients*

Trials involving several centres will inevitalby generate more difficulties with poor follow-up. In Britain, all patients randomised should therefore be « flagged » in the NHS central register, so that at least survival (and certified causes of death) are automatically completely followed up without further effort for the indefinite future. (For a fee of £1 per patient, a *bona fide* medical research worker who is running a trial in Britain can have his own professional address attached to the records of all his trial patients. When deaths occur he will be told the date of death, the certified causes of death, and the name of the doctor who certified the death. For details, write to OPCS, 10 Kingsway, London WC2. In the U.S.A., one « national death register » from 1.1.1979 onwards has been proposed by an NIH committee (G. Beebe, pers. comm.).

viii. *Keep a log*

If physicians (or their secretaries) are asked to keep a simple log of all patients who they ever see with the disease being studied, together with a note of *either* randomisation result *or* reason not randomised, then embarrassment at writing down false or inadequate reasons may increase the number randomised.

ix. *Telephone randomisation*

In large trials involving several centres slight uncertainty as to exactly who has been randomised may arise, so it is preferable to have all randomisations by telephone to a central office, giving full name, sex and exact date of birth. In all British trials, the statistician should use these data to « flag » each patient's name in Government archives, for this ensures complete survival follow-up even if the whole trial organisation later collapses. (Analysis can then be undertaken of survival from the date of telephoned randomisation; this is more objectively reliable than analysis from the report date of diagnosis, or first treatment.)

x. *Unstratified randomisation*

Stratified allocation is usually an unnecessary complication, as long as *retrospectively* stratified analysis is performed (see ref. 9, sections 12 and 22).

xi. *Unrestricted entry*

Relax all the « exclusion criteria » you can, and widen the class of patients being studied in any way you can — accept older patients, younger patients, patients with disease that is either more, or less, advanced than originally intended, etc. Having got a wide class of patients, any pure subset you like can easily be examined retrospectively, so no harm can ever be done by widening entry. (It is advisable, however, if the greatest effects are expected in one particular subgroup, to write out this expectation in the protocol, so that if the expected effect occurs people do not dismiss it as a chance interaction.)

xii. *National numbers*

Know the national incidence or death rate for the condition concerned, as this may be useful ammunition when you argue for larger trials.

« Mortality Statistics by Cause for 1975 » (HMSO publication, series DH2 N° 2) can be purchased for £2 + post (it weights 240 gms) from the Government Bookshop, 49 High Holborn, London, WCIV 6HB and is invaluable. For non-fatal conditions, the DHSS/OPCS publication « Report on Hospital In-Patient Enquiry for the year 1967, Part I, Tables » (HMSO 1970 £2) may be obtained from the same address, reporting hospitalisation by reason. Finally, for solid tumours, « Survival of Cancer Patients : Cases diagnosed in Norway 1953-67 » is useful as it splits them by « localised, regional or distant ». It may be obtained, usually free to *bona fide* research workers from the Secretary, Norwegian Cancer Society, Huitfeldsgt 49, Oslo 1, Norway.

xiii. *Give the maximal practicable dose*

You can increase the power of a trial either by randomising more patients or by having a larger treatment difference. If a treatment which works would probably work better if you gave more of it, then a trial of the maximal practicable dose against control would be more likely (with the limited numbers you will actually have) to yield a positive result than would comparison of a more moderate dose with control. Also, if even the maximal dose is ineffitive then probably the more moderate dose would also be ineffective, while the converse is not necessarily true. Trials studying extreme contrasts are therefore likely to be more informative.

xiv. *Randomise A vs. B1 vs. B2, but not A vs. B vs. C*

If collaboration in a trial of treatment A versus treatment B cannot be agreed easily because there is dispute as to which form of B (B1 or B2, say) should be tested (B1 and B2 might only differ in some lesser detail of a drug schedule, perhaps), then an acceptable compromise might be randomise equally between A, B1 and B2, with the eventual intention of comparing the one-third of patients who received A with the two-thirds who received either B1 or B2. (As a small bonus, a free trial of B1 versus B2 will have been done.) Apart from this, however, it is usually unwise to randomise between 3 very different treatments A, B and C, because your trial probably has hardly enough patients for a simple two-treatment comparison, let alone a 3-treatment comparison/on two « degrees of freedom » (unless similarities between A, B or C can be exploited to reduce the statistical analysis to a one-degree-of-freedom comparison).

xv. *Additional Studies*

A trial may be made more attractive to prospective collaborators by adding on biochemical, histological or clinical measurements which, although not essential for the success of the trial, are of independent scientific interest. However, care must be taken not to dissuade prospective collaborators or patients by increasing the complexity of the study, and, as has already been suggested, such studies may best be presented as optional extras. Trials may also be made more valuable scientifically by the use of factorial designs (see Appendix 2).

xvi. *Extend the period of intake*

If you plan that intake should last for (say) 2 years, then don't absolutely bind yourself to this in the protocol. If the original protocol merely says « It is expected that intake will continue for 2 years », then if intake is really going well during the second year and no other study is urgent and all ready to start, you can discuss with the participants the possibility of continuing intake (at several or all of the participating centres) for another year. Indefinite prolongation of intake is undesirable because interest, and hence intake, may fall off, but a definite extra year or two on a current study would often be more valuable than undertaking a completely new study.

APPENDIX 2

Factorial designs (particularly 2 × 2 designs)

For many trial organisers, this Appendix could prove to be the most immediately useful part of this paper, because « factorial » designs represent one of the few substantial improvements in clinical trial design which can be implemented with little or no extra difficulty or cost.

Suppose that we have two largely unrelated questions to answer concerning the treatment of the same disease. (This is a very common situation indeed, for there is usually no shortage of important questions.)

For example, suppose that for postoperative postmenopausal Stage II breast cancer patients we wish both to evaluate tamoxifen (an anti-oestrogen agent) and to evaluate cytotoxic therapy. Cytotoxic therapy is obviously not justified, especially for older women, unless it really helps, and even regular oral tamoxifen costs money, is a constant nuisance, its implication of residual disease may worry some patients, and it may turn out to have some unsuspected side effects (e.g. neuropathy, perhaps) on prolonged use. We already know, from previous trials, that the benefits of

either treatment are unlikely to be large for women in this age group, but they might still be sufficient to be medically important. As has already been discussed, to distinguish between a moderate benefit and a negligible benefit for either treatment a randomised trial involving perhaps 1000 or more women is desirable, a number which we can barely hope to achieve with our maximal effort. How should the treatments be allocated among our intake of (say) 900 women? Some investigators might decide to answer one question at a time (e.g. by putting 450 on cytotoxic treatment and having 450 controls), leaving the other question to other investigators or future studies, while some investigators might try to answer both questions by randomising 3 ways (e.g. 300 nil, 300 tamoxifen, 300 cytotoxic plus tamoxifen, or some other 3-way design). Both these are poor designs, the first because it needlessly fails to answer one of the important questions, and the second because each question is answered only by the comparison of 300 versus 300 instead of the full 450 versus 450 which is possible. A better design than either of these would be to use a « 2 × 2 » design*, in which patients are randomised four ways (225 nil, 225 tamoxifen only, 225 cytotoxic only, 225 both cytotoxic and tamoxifen).

To answer the question « Does tamoxifen help? » we would compare the 450 patients who received tamoxifen with the 450 patients who did not. (Half of each group would have received cytotoxic therapy, but that doesn't make the groups non-comparable — it could just be an extra factor to be retrospectively stratified for during analysis.) To answer the question « Does cytotoxic treatment help? », we could likewise compare the group of 450 patients who received cytotoxic therapy with the group of 450 who did not. In other words, by using a factorial design we have got two independent answers for the price of one. Surely such designs should be commoner than they now are? (There is no implicit assumption in 2 × 2 designs that tamoxifen is of *equal* value whether or not cytotoxic therapy has been given, merely the expectation that if tamoxifen is of *some* value for patients who have received cytotoxic therapy, then it is also likely to be of *some* value for patients who have not. If this expectation holds, 2 × 2 designs are efficient. If it does not hold, 2×2 designs will point unbiasedly to the complicated truth, while misleading conclusions could well emerge from other designs.)

The role of 2 × 2 designs could be very wide. For example, in Britain a working party of the Medical Research Council is currently responsible for a trial on 18000 people with moderate hypertension. The subjects will be randomised between hypotensive treatment and placebo, continued for 5-10 years, at a cost of over one million pounds, to see whether the incidence of vascular accidents is reduced. This sort of study would be ideal for a 2 × 2 design; a question such as « Is daily prophylactic aspirin advisable for people with no history of stroke or MI? » could be incorporated into such a study at little extra cost other than the provision of calendar packs of aspirin to 9000 people (4500 on hypotensive therapy and 4500 not). In view of the ability of one aspirin per day to inhibit platelet aggregation, and the apparent moderate effects of aspirin on mortality among myocardial infarction patients, the effects of aspirin on morbidity and mortality in an apparently healthy population urgently requires evaluation, and few better opportunities than this will arise. Indeed, one could imagine a 2 × 2 × 2 design for this trial, where a randomly chosen half of the patients are addition informed by letter of the likely relevance of salt to hypertension and are asked, with regular reminders, to try to avoid the addition of salt during the cooking or eating of their meals. If (which is, unfortunately, rather unlikely) a fair proportion of them did respond by cutting down their salt intake appreciably, a question which as been unanswered for a long time could perhaps be settled cheaply and directly.

I do not want to give the impression that 2 × 2 designs are only relevant to large studies, for they are equally advisable in all studies, large or small. Indeed, some questions (e.g. the role of vitamin C or laertrile in cancer therapy), which might not at present be thought to be of sufficient interest to justify setting up large trials just to answer them, could be answered by being included, in a 2 × 2 design, in several different trials which were primarily designed for other purposes.

A similar philosophy might be useful for treatments (such as asparaginase for remission induction in acute myeloid leukaemia) which were once thought pro-

* These designs were introduced half a century ago into agricultural research, and a quarter of a century ago into medical research. They were christened « 2 × 2 » designs because the allocation can be thought of as 2-way randomisation between tamoxifen versus nil, followed by 2-way randomisation between cytotoxic therapy versus nil. Clearly, just as one could design a trial with a 3-way allocation (A versus B versus C, with no similarities between A, B and C that can be exploited to produce the main comparison to one degree of freedom only) so also 3 × 2, or even 3 × 3, or larger, designs could be envisaged. All would be called « factorial » designs, but (unless there are relationships between the 3 treatments which could be exploited to reduce each statistical analysis to a one-degree-of freedom comparison) all would suffer from the usual power losses associated with 3-way randomisations. However, there is in principle no reason why 2 × 2 × 2 (8-way) factorial designs, or even, in large trials, 2 × 2 × 2 × 2 (16-way) factorial designs should not be considered to answer 3 or even 4 questions simultaneously and efficiently.

mising but which have now been tried and found not to be *very* effective. Even though further trials chiefly devoted to them might not be appropriate, they could still be included in several current trials by using a 2 × 2 design, and we could thereby discover, without much extra work, whether they confer a moderate or a negligible net benefit.

Let me finally recapitulate what I have said about 2 × 2 designs. Given two questions (e.g. treatment A or not? and, treatment B or not?) to answer, we randomise 4 ways (A and B, A only, B only, neither). Eventually, we shall evaluate A by using a retrospectively stratified analysis to compare all those allocated to A with all those not, and we shall evaluate B analogously. The only disadvantage of a 2 × 2 design would be if the complexity of it deterred patients or physicians from collaborating, or if one of combinations is contraindicated. If we know from the outset that we have two main questions to answer, a 2 × 2 design is probably very much better than any other. If we have only one main question to answer and we can organise a trial to answer it, then when the trial design is nearly settled we can sit back and ask what subsidiary questions would like a free answer to. If we have only one main question and we cannot organise a trial to answer it, then we can perhaps try to answer it parasitically by writing over the next few years to people who are planning or undertaking trials of other treatments to ask whether they will consider a 2 × 2 design that will answer our question as well as theirs. (One could both organise a trial oneself and try to parasitise other people's trials). Whether or not any of these options are actually exercised, they certainly should, as is too often not the case at present, at least be thought about.

REFERENCES

1. Byar D. P., Simon R. M., Friedewald W. T., Schlesselman J. J., De Mets D. L., Ellenberg J. H., Gail M. H. & Ware J. H. Randomised clinical trials : perspectives on some recent ideas. *New Engl. J. Med.*, 1976, *295*, 74.
2. Chalmers T. C., Block J. B. & Lee S. Controlled studies in clinical cancer research. *New Engl. J. Med.*, 1972, *287*, 75.
3. Chalmers T. C., Matta R. J., Smith H. & Kunzler A. M. Evidence favouring the use of anticoagulants in the hospital phase of acute myocardial infarction. *New Engl. J. Med.*, 1977, *297*, 1091.
4. Hill J. D., Hampton J. R. & Mitchell J. R. A. A randomised trial of home-versus-hospital management for patients with suspected myocardial infarction. *Lancet*, 1978, *1*, 837.
5. Mather H. G., Morgan D. C., Pearson N. G., Read K. L. Q., Shaw D. B., Steed G. R., Thorne M. G., Lawrence C. · & Riley I. S. Myocardial infarction : a comparison betw .en home and hospital care for patients. *Brit. med. J.*, 1976, *1*, 925.
6. McPherson K. Statistics : the problem of examining accumulating data more than once. *New Engl. J. Med.*, 1974, *290*, 501.
7. Medical Research Council. Responsibility in investigations in human subjects. *Report for the year 1962-3*, 1964, 21.
8. Moertel C. G. & Reitemeier R. J. Advanced Gastrointestinal Cancer/Clinical Management and Chemotherapy. Hoeberl Medical Division, *Harper and Row*, New York, 1969.
9. Peto R., Pike M. C., Armitage P., Breslow N. E., Cox D. R., Howard S. V., Mantel N., McPherson K., Peto J. & Smith P. G. Design and analysis of randomised clinical trials requiring prolonged observation of each patient. *Brit. J. Cancer*, 1976, *34*, 585 and 1977, *35*, 1.
10. Pocock S. J. Randomised clinical trials (letter). *Brit. Med. J.*, 1977, *1*, 1661.