STATISTICS IN MEDICINE, VOL. 3, 409–420 (1984)

WHY DO WE NEED SOME LARGE, SIMPLE RANDOMIZED TRIALS?

SALIM YUSUF* RORY COLLINS AND RICHARD PETO
Clinical Trial Service Unit, Radcliffe Infirmary, Oxford, UK

The criteria for a good trial are similar in many serious diseases: first and foremost, ask an 'important' question and, secondly, answer it 'reliably'. These two very general criteria obviously require further elaboration, but even as they stand they can suggest some surprisingly specific consequences for clinical trial design. Particularly, they can be used to suggest both the possibility and the desirability of *large, simple randomized* trials of the effects on *mortality* of various *widely practicable* treatments for *common* conditions.

There are six main steps in the argument leading to this conclusion. First, the identification of effective treatments is likely to be more 'important' if the disease to be studied is *common* than if it is rare, and studies of common conditions *can* be large.

Secondly, the identification of effective treatments for common diseases is likely to be more 'important' if the treatment is widely practicable than if it is so complex that it can be performed only in specialist centres, and treatment protocols for widely practicable treatments *can* be simple.

Thirdly, study of the effects of treatment on major endpoints (e.g. death) is likely to be more 'important' than study of its effects on minor endpoints (e.g. radiological or biochemical evidence of disease recurrence or progression that the patient is not directly aware of), and follow-up protocols for the assessment only of major endpoints *can* often be simple.

Fourthly, no matter what prognostic features are recorded at entry, the duration of survival, etc. among apparently similar patients is likely still to be rather unpredictable, so no great increase in statistical sensitivity is likely to be conferred by stratification and/or adjustment for such features. In other words, the reliability of the main treatment comparison is improved surprisingly little by adjustment for any initial imbalances in prognostic features, which suggests that entry protocols too *can* be simple.

Fifthly, for reasons that will be discussed later, the direction (though not necessarily the magnitude) of the net effects of treatment on particular modes of death is likely to be similar in many different subcategories of randomized patient. Again, therefore, if there is no need to subcategorize patients to decide who needs which treatment, the entry protocol *can* be simple.

These first five points suggest that many 'important' trials *can* be large, and *can* have simple protocols. The sixth point is that for a question to be important it must not yet have been answered reliably, and that if a widely practicable treatment had a large effect on an important endpoint (e.g. mortality) in a common disease this would probably already be known, so the true effect is likely to be either null or, at most, moderate. This sixth point (the *moderateness* of what can plausibly be hoped for) has perhaps the most profound implications of all, for although the currently usual sort

* Present Address: Clinical Trials Branch, National Heart, Lung & Blood Institute, Bethesda, Maryland 20205, U.S.A.

0277–6715/84/040409–12$01.20

of trial size may suffice to detect a large treatment effect, it may well fail to detect a moderate (yet worthwhile) effect. Likewise, although the moderate biases that can be introduced by sloppy methodology (non-randomized controls, failure to use an 'intention-to-treat' analysis, undue emphasis on data-derived subsets of one trial, or on one or two particular trials rather than on the totality of the evidence from all relevant trials) may not matter much in those rare instances when large treatment effects await discovery, they matter enormously when only moderate effects are to be assessed.

Methodology that may introduce moderately large biases or moderately large standard errors is not appropriate for the assessment of the type of moderate treatment effects that are all that it is usually plausible to hope for. It is chiefly because one usually needs to be able to distinguish reliably between moderate and null effects that trials need to be *strictly* randomized, analysed and interpreted completely unbiasedly, and much, much larger than is currently usual. There are three practical implications of all this.

(1) For the realistic assessment of the effects of today's widely practicable treatments on mortality from the currently common neoplastic diseases or the currently common vascular diseases, the use of 'historical controls', 'databases', or whatever, is of little real value, for such methods may introduce *moderate* biases—and, although moderate biases may not matter much when extremely large treatment effects await discovery, they may matter considerably when only moderate effects are to be assessed.

(2) Moreover, most of today's randomized trials are too small to be of much independent value (except as contributors to a formal overview of many randomized trials of one question: see below). The practical problems of producing tenfold increases in trial size can, however, be greatly eased by remembering that trials of the effects of widely practicable treatments on major endpoints in common conditions *can* be ultra-simple, and hence large. Indeed, one can probably go further, and say that a number of important medical questions will in the next few years be answered reliably only if some ultra-simple, ultra-large, strictly randomized trials can be mounted.

(3) Both in order to minimize random errors *and in order to avoid bias* when trying to assess a particular treatment, inference should wherever possible be based not on just one, or a few, of the available trial results, but on a systematic overview of the totality of the evidence from all the randomized trials.

This general perspective on clinical trials, however, rests on the two basic assumptions (the likelihood that the treatment effects to be assessed are at best only moderate, and the unlikelihood of unanticipated qualitative interactions[1]) that are matters of judgement, and it is therefore important to consider carefully the circumstances in which these assumptions are likely to be appropriate.

WHY ARE MODERATE EFFECTS ON MAJOR ENDPOINTS GENERALLY MORE PLAUSIBLE THAN LARGE ONES?

By implication, an 'important' question is one that has not yet been answered reliably. If any widely practicable intervention had a very large effect, e.g. a cure for most patients with a previously untreatable common condition, then whether or not randomized trials are desirable, these huge gains in therapy are likely to be identified more or less reliably by simple clinical observation, by 'historically controlled' comparisons, or by a variety of other informal or semi-formal non-randomized methods that may well suffer from moderate biases, but that will eventually yield a reliable consensus. Although large therapeutic improvements may be accepted more rapidly if the

stringent test of a randomized controlled study is undertaken, they will probably eventually gain acceptance anyway. So, if there remains some controversy about the efficacy of any widely practicable treatment, its effects on major endpoints may well be either nil, or moderate, rather than large.

This does not, of course, apply to its effects on various less fundamental endpoints, e.g. various measures of the extent or progress of the disease. For example, in patients with myocardial infarction it is not difficult to demonstrate that ventricular arrhythmia may be reduced with lidocaine, infarct size limited with beta-blockers and coronary thrombus lysed with streptokinase, whereas for cancer patients it may be fairly easy temporarily to shrink a tumour. In each of these cases however, it is extremely difficult to prevent (or substantially delay) a large proportion of *deaths*. Indirect support for this conclusion comes from many sources, including (a) the previous few decades of disappointingly slow progress in the curative treatment of the common chronic diseases of middle and old age; (b) the heterogeneity of each single disease, as evidenced by the unpredictability of survival duration even when apparently similar patients are compared with each other; (c) the variety of different mechanisms in some diseases (e.g. myocardial infarction) that can in principle lead to death, only one of which may be appreciably influenced by one particular therapy; and (d) experience with many earlier trials, review of which (Table I) suggests that the true risk reductions being studied were probably only of the order of 5, 15 or 25 per cent, rather than (for example) 40, 50 or 60 per cent.

Points (b), (c) and (d) perhaps deserve some expansion and illustration. First, although several important prognostic features can be identified in patients with a disease such as myocardial infarction, the *exact* mechanism(s) by which they are related to a *particular* patient's outcome is unclear. For example, we often do not know why one patient with a particular complication such as ventricular tachycardia may live for several years whereas another may die tomorrow. In other words, even *within* subgroups of patients who are, by currently available criteria, fairly similar to each other, there will be considerable heterogeneity of outcome, i.e. the outcome may chiefly be

Table I. 'Plausible' risk reductions in the treatment of patients with myocardial infarction

	No. of trials	Estimated mortality reduction (% & SE)
A. *From acute treatment*		
I.V. Thombolysis	19	22% ± 5%
Glucose–insulin–potassium	*5	*23% ± 12%
I.V. Nitrates	*6	*30% ± 14%
Hyaluronidase	*5	*36% ± 15%
Oral beta-blockade	22	7% ± 12%
I.V. beta-blockade	27	8% ± 14%
B. *From long-term treatment*		
Aspirin	6	10% ± 6%†
Sulfinpyrazone	*2	*15% ± 12%
Anticoagulants	*10	*20% ± 7%
Beta-blockade	24	22% ± 4%

I.V. = Intravenous. S.E. = Standard Error.
These data are based on pooled estimates, most of which are not yet published, from all available trials of these agents using the technique of retrospective stratification.
* Not necessarily trustworthy, as some trial results may not yet have been traced.
† 12 ± 6 per cent, if chance imbalances in initial prognostic features are allowed for: see Reference 2.

dictated by unknown or unmeasured features (and hence, perhaps, by mechanisms that we do not deliberately affect).

Secondly, in many chronic diseases, e.g. coronary artery disease, we may know of several different mechanisms that could lead to death, and current drug treatments are usually aimed at modifying only one or two of these at a time. (In a sense, many of the intervention trials—e.g. those involving antiplatelet or lipid-lowering therapy—have been as much a test of the relevance of a particular mechanistic hypothesis as they have been tests of the efficacy of a particular drug.) For example, elevated catecholamines, elevated free fatty acids, increased platelet aggregation, coronary artery narrowing and occlusion, cardiac muscle wall stress, etc., are each associated with somewhat different mechanisms by which prognosis in a myocardial infarct could be worsened. Although these mechanisms are unlikely all to be of equal importance, there are so many of them that it might be unrealistic to expect modification of any single one factor by a particular intervention to reduce mortality by a *large* proportion (e.g. by more than 50 per cent) although it may be realistic to hope for some more *moderate* benefits (e.g. reducing mortality by 10 or 20 per cent). Finally, returning to the rather sobering data in Table I, in each case the mortality reduction suggested by the pooled data is only moderate (e.g. 10, 20 or 30 per cent). Indeed, for the more reliably evaluated therapies such as intravenous fibrinolytic drugs, long-term beta-blockers and long-term aspirin, the mortality reductions appear rather modest (22, 22 and 10 per cent, respectively).

Having accepted that, with many currently available agents, only *moderate* reductions in mortality are plausible, how worth while might such effects be, if they could be reliably detected? To some clinicians, reducing the risk of death in patients with myocardial infarction from 10 per 100 patients to 8 or 9 per 100 patients treated may not seem particularly worth while, and, indeed, if such a reduction was achievable only at the expense of prolonged treatment by an expensive or toxic agent, this might well be an appropriate view. On the other hand, since death from myocardial infarction is common, a simple, non-toxic, widely practicable treatment that reduced the risk of death by perhaps 10 or 20 per cent could, on a national (or international) scale, have substantial public health implications and might, for example, prevent or substantially delay several tens of thousands of deaths a year in the U.S. alone. Many of the patients who would benefit would, moreover, still be in middle, rather than old, age, with a reasonable expectation of enjoyable life. These absolute gains are substantial (and might, indeed, considerably exceed the numbers of lives that could hypothetically be saved by a simple cure for all patients with some less common disease such as acute myeloid leukaemia).

THE NEED FOR LARGE TRIALS

Having accepted that mortality reductions of perhaps 15 per cent or so are all that is plausible for many currently available treatments, and that such modest reductions are worth detecting, one aim of current and future trials should be to distinguish reliably between the only two medically plausible alternatives: either there is *no* worthwhile difference in survival, or treatment confers a moderate, but worthwhile, benefit (e.g 15 or 20 per cent fewer deaths). It is not sufficiently widely appreciated just how large clinical trials really need to be, in order to detect such moderate differences reliably. This can be illustrated by a hypothetical trial that is actually quite inadequate, even though by current standards it is moderately large, in which a 20 per cent reduction in mortality is supposed to be detected among 2000 patients (1000 treated and 1000 not). If this were the case, then one might predict about 100 deaths (10 per cent) in the control group and 80 (80 per cent) in the treated group. Even if exactly this difference were observed, however, it would not be conventionally significant ($P = 0{\cdot}1$), and although the play of chance might well increase the

difference enough to make it conventionally significant (e.g. to 110 deaths vs 70 deaths; $P < 0{\cdot}001$), it might equally well dilute, obliterate or even reverse it (e.g. to 90 deaths vs 90 deaths). In real life, of course, the situation is even worse than this, as the average trial size is probably nearer to 200 then to 2000 patients!

The unsatisfactory nature of the present situation is nicely illustrated by the fact that, of the 24 trials that have attempted to evaluate long-term beta-blockade,[2] 21 have failed to achieve conventional levels of statistical significance, mostly because they were too small to be reliable. In aggregate, the results of these 24 trials suggest a mortality reduction of 22 per cent (with standard error only ± 4 per cent, and so undoubtedly real). But, even if the true risk reduction were about a quarter, most of these 24 trials would still probably have missed it. Table II summarizes the *expected* relationship between trial size and trial outcome for such studies, and Table III summarizes the *actual* relationship. These tables deserve detailed scrutiny, for their implications are important and disturbing. (Moreover, the effects on mortality of beta-blockers may well be somewhat *larger* than those of most other currently available promising agents that await evaluation!) Reliable monitoring of 15 per cent risk reductions might, for some agents (e.g. long-term low-dose enteric-coated aspirin), be medically worthwhile, yet would require trials substantially larger than the largest of the beta-blocker trials. Indeed, reliable assessment of such important differences generally requires not trials in which one or two thousand patients are entered, but instead trials in which one two thousand *endpoints* occur. This is likely to involve the randomization of some 10,000 or 20,000 patients with a good prognosis or of several thousand patients with a poor prognosis. The evolution of collaborative groups that will result in significant questions being addressed by such trials at a practicable cost is one of the major challenges now facing clinical research.

CAN SIMPLE TRIALS PROVIDE RELEVANT AND RELIABLE ANSWERS?

A key principle underlying the argument that clinical trials can be simple and yet provide medically relevant conclusions involves careful distinction between 'quantitative' interactions and 'qualitative' interactions.[1] In this context, a qualitative interaction is one whereby the true treatment effects in different subgroups (e.g. male and female) do not even point in the same direction, whereas a quantitative interaction is one whereby the direction of the true effect of treatment is similar, and only the size of the benefit is different. Interactions of any non-zero treatment effect with almost any prognostic variable are commonly to be expected, but these are likely to be *quantitative* rather than *qualitative*. For example, in a trial of an anti-arrhythmic agent, one might expect among patients who appear to have the worst arrhythmias (e.g. rapid and prolonged ventricular tachycardia) a risk reduction different from that among those who appear to have less severe arrhythmias, but presumably one would expect some effect in both categories of patient. Such 'quantitative' differences in the size of the treatment effect in different subgroups are quite likely to exist even if formal statistical tests for interactions are not conventionally significant. By contrast, *unanticipated qualitative* interactions (whereby treatment is of substantial benefit among one recognizable category of patients in a trial and not among another) are probably extremely rare, even though in retrospective subgroup analyses they may seem extremely common. Of course, one can recognize *a priori* certain categories of patients for whom certain drugs are contra-indicated (e.g. for patients with severe heart failure or advanced heart block, a beta-blocker is so clearly contra-indicated that such patients would probably have been formally ineligible for a beta-blocker trial). Our expectation is not that *all* qualitative interactions are unlikely, but merely that *unanticipated* qualitative interactions are unlikely, especially if attention is restricted to one mode of death.

Table II. Expected effects of trial size on trial results. Relationship between the total number of deaths* in the two treatment groups combined and the approximate probability of a convincingly ($1P < 0{\cdot}01$) significant result emerging, in trials where allocation to active treatment actually reduces the odds of death by about a quarter,† which is an effect sufficiently large to be worth knowing about.‡

Total no. of deaths* (treated + control)	(Approx. no. of patients randomized if risk $\simeq$ 10 per cent)	Approx. probability of failing to achieve $1P < 0{\cdot}01$ significance if true risk reduction $\simeq 1/4$	Comments that might be made on sample size before trial begins
0–50	(under 500)	over 0·9	Utterly inadequate
50–150	(1000)	0·7–0·9	Probably inadequate
150–350	(3000)	0·3–0·7	Possibly adequate, possibly not
350–650	(6000)	0·1–0·3	Probably adequate
over 650	(10,000)	under 0·1	Definitely adequate

* About twice as many patients would be needed to achieve corresponding probabilities of detection of risk reductions of only 1/6 (instead of 1/4). Conversely, only about half as many patients might be needed for risk reductions as large as 1/3.

† Because of the inevitable diluting effects of non-compliance and certain other protocol deviations, a 25 per cent risk reduction might be produced by allocation to a treatment that actually confers a risk reduction of about 30 per cent.

‡ Although it would be *nice* to be studying a treatment that really halved the risk of death (both because of its practical value and because trials on a few thousand patients are much easier to undertake than trials on several thousand patients are), experience suggests that few, if any, of the widely practicable treatments that currently require evaluation by clinical trials in any of the common life-threatening diseases are likely to confer such large benefits. Moreover, if some widely practicable treatment for a common life-threatening disease did produce a reduction of 'only' 25 per cent or so in the odds of death, which (whatever one may hope for at the inception of a trial) is often about as much as is reasonably likely to be the case, this might well be medically worthwhile for many patients, so one would wish not to miss such an effect. Hence, trialists should ideally try to make their trials large enough to detect reliably the kind of worthwhile but moderate treatment effects (reductions by 1/6 or 1/4 in the odds of death) that it is medically realistic to hope for, as long as pursuit of this ideal does not, in cases where it appears impracticable, result in discouragement of the prospective investigators and abandonment of plans for a trial that, although inadequate on its own, could have contributed usefully to an overview of many tirals, as in this paper.

Table III. Actual effects of trial size on trial results. Relationship between the total number of deaths in the two treatment groups and the result actually attained, in the 24 trials of a treatment (long-term beta-blockade) that reduces the odds of death by about* 22 per cent (with SE ± 4 per cent: Table I).

Total no. of deaths in trial (β-bl. ± plac.)	(Mean no. of patients randomized)	Statistical power (see Table II)	No. of trials resulting in:			
			$P < 0{\cdot}05$ against	Non-sigt. against	Non-sigt. favourable†	$P < 0{\cdot}05$ favourable
0–50	(255)	Utterly inadequate	0	5	5	0
50–150	(861)	Probably inadequate	0	1	9	1
150–350	(2925)	Possibly adequate, possibly not	0	0	1	2
350–650	(No such β-bl. trials exist)	Probably adequate	—	—	—	—
over 650	(No such β-bl. trials exist)	Definitely adequate	—	—	—	—
Total	(866)	Inadequate separately, adequate only in aggregate	0	6	15	3

* The pooled percentage reduction in the odds of death, based on the overview of these trials in Table I, is 22 ± 4.
† Includes one very small trial with zero difference.

These expectations suggest that in a hypothetical extremely large trial of an active treatment the results in different subgroups would tend to point in the same direction. However, in a few particular subgroups of an actual trial this tendency may be either substantially exaggerated or, conversely, diluted or reversed by the play of chance. Likewise, although the results in several extremely large trials of a particular treatment should in theory all be similar, many trials are not large, and so the play of chance may actually produce some apparent heterogeneity in the direction of the effects of similar treatments in different trials. In this context, it is self-evident that many of the pitfalls associated with the selective interpretation of preferential treatment effects in particular subgroups *within* a trial are equally applicable only if a few of the relevant trial results (selected, perhaps, to be the most or least promising) of a particular agent are unduly emphasized. Consequently, not only should the conclusions in an individual trial be based on its overall result, but also, as long as bias can be avoided, a semi-formal overview of all the relevant trials (which minimizes the effects of the play of chance) may be considerably more informative than any one trial. This in turn suggests the wider generalizability of any effects that are clearly evident in such overviews of trials to types of patients other than those who were entered (and to subgroups where no apparent benefit was observed) as long as there are no strong prior reasons to expect harm, i.e. no strong anticipation of a qualitative interaction. Therefore, much less detail per patient than is usual is all that is required to assess the efficacy of a treatment validly. Such simplicity may in turn facilitate the attainment of the trial sizes (e.g. some tens of thousands of cardiovascular patients) that need to become commonplace if randomized trials are really to fulfil their enormous promise, for a tenfold increase in trial size probably requires a comparable decrease in costs per patient.

HOW CAN SIMPLICITY BE ACHIEVED IN TRIALS WITHOUT BIAS?

Typically, trials need to be capable of distinguishing reliably between the two initially plausible alternatives: either the treatment confers no material benefit, or it has a worthwhile, but moderate (e.g. 15 per cent or so) effect on mortality. Therefore, both the random and the systematic errors inherent in whatever trial methodology is to be used to assess such treatments should be small in comparison with a 15 per cent mortality reduction. This requires strict control of both *systematic* and *random* errors.

Systematic error (i.e. bias) can be avoided by randomization followed by an unbiased statistical analyis by *allocated* rather than by *actual* treatment, and by basing inference chiefly on the results in the entire trial (or, preferably, on an overview of the results of all relevant trials) and not on some *post hoc*, data-dependent subset of the patients in the trial (or on only some, but not all, of the relevant trials). The argument for randomization is not that no truths can emerge without it—indeed, the history of medicine contains many examples where uncontrolled clinical observation has reliably established the value of certain treatments—but that without it *moderate* biases can easily emerge.

Random error can be minimized only by achieving a really large trial size—indeed, if 15 per cent mortality reductions are to be assessed then there may need to be over 1000 endpoints available for treatment evaluation in the aggregate of all relevant trials. Precise characterization of prognostic features at entry is, perhaps surprisingly, unlikely to confer any material reduction in the number of endpoints that need in aggregate to be studied. (For example, if mortality in an MI trial was 1000/10,000 (10 per cent) non-diabetics and 200/1000 (20 per cent) diabetics, knowledge of the diabetic status at entry would increase the effective sample size by less than 1 per cent. Indeed, even in the Aspirin Myocardial Infarction Study,[4] a large trial where, at considerable expense, many dozens of important prognostic features were recorded, the increase in the effective sample size that this mass of extra data conferred on the treatment comparison was only about 10 per cent!)

The positive virtues of simplicity of trial design probably become fully apparent only when the need for a really large randomized trial is properly understood. Many clinicians involved in the management of patients are already overworked, and in practice a really large-scale trial is likely to succeed only if it adds little or nothing to their existing workload. Ideally, not only should the extra work be reduced to almost nothing (simple trial entry and patient management, few or no extra follow-ups, very few forms—and these extremely simple—etc., etc.), but also the clinician should be given an amount of extra nursing, secretarial or other help per trial patient that clearly outweighs whatever extra work the trial does involve. Ideally, it should be less trouble to collaborate wholeheartedly in a trial than not to do so, for many hundreds of medical collaborators may have to be involved to some extent in a really large trial.

Various measures may help achieve maximal simplicity. First, of course, one may deliberately choose to avoid ancillary studies or expensive investigations, for these may impede rather than assist reliable assessment of the effects of treatment on mortality. Secondly, one may make entry criteria as wide and as simple as possible; *they may even vary substantially* from centre to centre* (within broad guidelines). Thirdly, one may use telephone randomization first obtaining by phone the patient identifiers, the *few* baseline variables that are known to be prognostic importance and the *few pre*-selected factors that may be particularly likely to 'interact' strongly with treatment. Telephone randomization has the advantage of yielding complete entry data without troubling participants for any entry form, and, moreover, it yields a reliable, precise list of every patient ever randomized.

The treatment regimen must be made as simple as possible, since complex schedules may result in poor investigator and patient compliance. Moreover, if the results of the trial are favourable, simple treatments are more likely to find widespread applicability and acceptability. To encourage pill taking, calendar packs can be used, but in general, while encouraging compliance to the fullest, it needs to be measured only crudely by simple techniques such as patient reporting or pill counting. (In many studies, except those of behaviour modification, the accuracy of pill counting or patient reporting does not appear to be materially different from that of more sophisticated biochemical tests—and, it really does not affect the interpretation of the trial if the estimated compliance was 85 per cent instead of 80 per cent.) The complexity and number of follow-up forms should be minimized; for example, it may be advisable to avoid going through a systematic list of any known minor side-effects (by definition, we already know about them!) and to ask about only those few major side-effects that would cause medication to be stopped or treatment schedules to be altered. One practical method may simply be to ask the patients if they experienced any side-effects that worried them, and to ask them for a list of all medical conditions that 'made them seek medical care', without using detailed questionnaires. Such an approach would have the advantage of detecting serious *unexpected* adverse effects, which a detailed questionnaire of *known* side-effects might miss. Measurement of death as an endpoint is simple and carries no bias as long as the analysis is carried out chiefly by allocated, not actual, treatment groups. In some countries such as the U.K., Sweden, Norway, Denmark, Finland, the U.S. and Canada, cause-specific mortality can be reliably obtained from centralized government archives for a nominal cost, at little extra effort.

---

* Even if strict entry criteria were devised, there is likely to be substantial unmeasured heterogeneity among apparently similar patients. Moreover, different physicians may apply nominally similar entry criteria rather differently—for example, the exact time of onset of infarction may be variably interpreted, the definition of 'severe' chest pain or heart failure may be different, etc. Thus, even patients chosen by similar 'objective' criteria, e.g. right bundle branch block following MI, may actually be very heterogeneous as some may have a variety of additional unrecorded features that may cause large variations in outcome. Further, there is rarely a need for absolute homogeneity (see earlier discussion on quantitative and qualitative interactions) in a clinical trial; what is needed is merely lack of bias between the two treatment groups.

(In others, such as France, Belgium or Italy, dates but not causes at death can be obtained.)

A final way of reducing the cost of evaluating treatments is to get two answers for the price of one by the use of a 'factorial' (e.g. 2 × 2) trial design. Indeed, wherever feasible, factorial designs should be considered. It is commonly, but mistakenly, believed that factorial designs are appropriate only in the absence of interaction, but this is not at all the case: all that is needed is the lack of any *qualitative* interactions between the effects on the treatments that are to be studied. (Moreover, in the unlikely event that some qualitative interaction does exist, then a factorial trial may well be the design of choice, since it alone will point unbiasedly to the complicated truth.)

TWO EXAMPLES OF LARGE, INEXPENSIVE, SIMPLE RANDOMIZED TRIALS

A. An acute intervention trial—ISIS

ISIS (International Study of Infarct Survival) is a uniquely large and uniquely simple trial. The study is a randomized, controlled study of the effects of early intravenous beta-blockade on mortality. Well over 250 centres in 14 countries randomize patients by a 2-minute telephone call to a centralized 24-hour randomization service. All entry data (heart rate, blood pressure, age, sex, diabetes, previous MI), other than the entry ECG, are collected by telephone—there is no entry form—and random allocation to open treatment or open control is then given in response. This provides 100 per cent entry data, no form-filling and complete information at the central co-ordinating office of exactly who was randomized. The principal analyses will be of 7-day and subsequent cardiovascular mortality among all randomized patients, and will compare all those allocated treatment with all those allocated control. The treatment regimen is straightforward, and consists of a 10–15 minute intravenous injection of a beta-blocker (atenolol) followed by oral atenolol once daily for 7 days. At discharge, a simple 1-page data form and the pre-randomization ECG are returned to the co-ordinating office. The long-term follow-up is almost entirely restricted to mortality, and is mostly done by the co-ordinating centre (though central government archives), so that no further work is generally required by the clinician entering the patients.

This experiment in trial design may well succeed, as the intake rate has been 400–600 patients per month (total: over 16,000 patients by closure at the end of 1984). The trial will have good power to detect a 15 per cent reduction in mortality. Of course, there are many interesting questions that ISIS will fail to address, for no single experiment can be expected to answer everything. One of the chief such questions will, as always, be the perennial problem of identifying exactly who needs treatment and who does not, except so far as this is determined by the recorded factors of age, sex, history of MI or diabetes, or pre-randomization ECG, heart rate or blood pressure. But, in exchange for failing to address various other questions, this trial will provide uniquely reliable evidence about the average effects of the widespread use of such treatment on mortality. It will, moreover, do so at a practicable cost (of about two million U.S. dollars for just over 16,000 patients) and is already the largest secondary prevention trial that has ever been done.

B. A long-term trial—the UK-TIA trial

Ordinary aspirin has profound effects on platelet function, and for a few days following ingestion of just one standard (300 mg) aspirin tablet platelet aggregability is grossly impaired and the bleeding time is moderately increased. Transient ischaemic attacks (TIAs) themselves are, by definition, of little clinical importance, but patients who have had one or more TIAs are at considerably increased danger of subsequent disabling stroke or myocardial infarction, the annual risk being of the order

of 5 per cent. Obviously, it would be nice if this risk could be reduced substantially (e.g. to 2 or 3 per cent). The disease is so common, however, and daily aspirin is such a simple treatment, that even a moderate reduction (e.g. to 4 per cent) might be of substantial practical importance. Unfortunately, no one neurologist or small group of neurologists would see enough cases to randomize the few thousand TIA patients that might be needed to estimate reliably such a moderate effect. Despite the absence of a tradition of collaboration among British neurologists in the 1970s, since 1979 some 50 consultant neurologists—about one-third of the entire profession—have been collaborating in a Medical Research Council financed, placebo-controlled study of daily aspirin. This trial, which is recruiting some 500 patients each year, is already (at over 2000 patients) bigger than all six other such trials in the world put together, and may eventually include nearly 3000 patients. One of the keys to its success may be its simplicity. Patients are randomized by direct-line telephone to the Clinical Trial Service Unit in Oxford, where name, sex, hospital and date of birth are recorded, and an individual two-letter treatment code is issued (e.g. DH). After this telephone call the patient is irrevocably in the study, unless it eventually appears that the symptoms were due to cancer in the brain. The neurologist then merely sends Oxford a simple single-sided notification form—and 4-monthly single-sided follow-ups—and writes out regular prescriptions for 'UK–TIA study pills, code DH' (for the pharmacist to sort out from the secret code lists). Urine tests have shown that errors are extremely rare and that compliance is good (and accurately self-described, so we didn't need the urine tests!). Collaborators are given one-fifth of a secretary or one-tenth of a doctor per 20 trial patients, in the hope of making it less trouble to collaborate in the trial than not to do so. If recruitment continues for the next couple of years, the main therapeutic question (about prevention of major vascular accidents) should be answered quite reliably. Whether this will happen or not depends entirely on the collaborators, so the trial results belong to them, not the co-ordinators. The final 7-year cost of the trial is likely to be about $1·5 million, chiefly for the day-marked, calendar-packed tablets. Even in Britain alone this is considerably less than the total cost of widespread use of aspirin if it is actually ineffective, or the cost of widespread non-use of aspirin if it does reduce the annual risk of serious vascular accidents from 5 to 4 per cent—and, the British trial results should also influence practice in many other countries.

HOW RELEVANT ARE RESULTS FROM LARGE, ANONYMOUS TRIALS TO REAL CLINICAL PRACTICE

Clinicians are used to dealing with individual patients, and may feel that the results of large trials somehow deny the individuality of each patient. This is almost the opposite of the truth, for one of the main reasons why trials have to be large is just because patients are so different from one another. Two apparently similar patients may run entirely different clinical courses, one remaining stable and the other progressing rapidly to severe disability or early death. Consequently, it is only when really large groups of patients are compared that the proportions of truly good and bad prognosis patients in each can be relied on to be reasonably similar. One commonly hears statements such as: 'If a treatment effect isn't obvious in a couple of hundred patients then it isn't worth knowing about'. Such statements often reveal not clinical wisdom but statistical naïvity.

One also hears it said that what is really wanted is not a blanket recommendation for everybody, but rather some means of identifying those few individuals who really stand to benefit from therapy. If any criteria (e.g. short-term response to a non-placebo-controlled course of some disease-modifying agent) can be proposed that are likely to discriminate between people who will and who will not benefit, then these can, of course, be recorded at entry and the eventual trial results subdivided with respect to them. There is, however, a danger in too detailed an analysis of the apparent response of small subgroups chosen for separate emphasis *because* of the apparently

remarkable effects of treatment in those subgroups. For, even if an agent brought no benefit, it would have to be acutely poisonous for it not to appear beneficial in one or two such subgroups! The surprising extent to which this applies is discussed, with a numerical example, in Reference 1. A large, anonymous trial will at least still help answer the practical question of whether on average a policy of widespread treatment (except where clearly contra-indicated) is preferable to a policy of no treatment (except where clearly indicated). Moreover, without a few really large trials it is difficult to see how else such questions could be resolved over the next few years. Digitalis, for example, has already been in use for over two centuries, and there is still no reliable consensus as to its net long-term effects. Trials are at least a practical way of making some solid progress, and it would be unfortunate if desire for the perfect (i.e. knowledge of exactly who will benefit from treatment) were to become the enemy of the possible (i.e. knowledge of the direction and approximate size of the effects of the treatment of wide categories of patient).

CONCLUSION

The present emphasis on large, simple trials is merely because the need for such trials has not yet been sufficiently widely recognized, and not because all other types of research are inappropriate. Obviously, a wide spectrum of research is needed, including many past, current and future trials quite different in concept from those described above. The intent of this article is not to suggest that all trials should be designed in one particular way, but merely to stress the need for *some* very large, very simple trials of widely practicable treatments. The more widely practicable the treatment, the simpler the treatment protocol can be, and the more serious the endpoint, the simpler its assessment may be.

REFERENCES

1. Peto, R. 'Statistics of cancer trials', in Halnan, K. E. (ed.), *Treatment of Cancer*, Chapman & Hall, London, 1981.
2. Editorial. 'Aspirin after myocardial infarction', *Lancet*, **1**, 1172–1173 (1980).
3. Yusuf, S., Peto, R., Lewis, J., Collins, R. and Sleight, P. 'Beta-blockade during and after myocardial infarction: an overview of the randomised trials', *Progress in Cardiovascular Disease* (in press) (1985).
4. Aspirin Myocardial Infarction Research Group. 'Randomised controlled trial of aspirin in persons recovered from myocardical infarction', *Journal of the American Medical Association*, **243**, 661–669 (1980).
5. Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J. and Smith, P. G. 'Design and analysis of randomised clinical trials requiring prolonged observation of each patient. I: Introduction and Design', *British Journal of Cancer*, **34**, 585–612 (1976).