CHAPTER 4
The Justification for Randomized Controlled Trials

The concept of random allocation when comparing different treatments has been an important aspect of the design of scientific experiments ever since the pioneering work of Fisher (1935). The first randomized experiments were in agriculture where the experimental units were plots of land to which the treatments, various crops and fertilizers, were assigned in a random arrangement. The purposes of such randomization were:

1. To guard against any use of judgement or systematic arrangements leading to one treatment getting plots with poorer soil, i.e. to avoid bias.
2. To provide a basis for the standard methods of statistical analysis such as significance tests.

In most types of non-human experiment, the investigator has all his experimental units available at once and can maintain tight control over how the experiment is conducted so that randomization can usually be implemented with only minor inconvenience.

However, the situation is very different for a clinical trial, in which the experimental units are patients. The idea that patients should be randomly assigned to one or other form of treatment is not intuitively appealing either to the medical profession or the layman. Superficially, randomized comparison of treatments appears contrary to the need for the clinician to give every patient the best possible care and hence appears to imply a loss of freedom for both patient and clinician. So, why should randomization now be considered such a mechanical detail of how randomization is actually prepared and carried out? The ethical and practical issues associated with randomization will be discussed. The mechanical details of how randomization is actually prepared and carried out will be explained in chapter 5.

4.1 PROBLEMS WITH UNCONTROLLED TRIALS

Traditionally, medical practice entails the doctor prescribing for a patient that treatment which in his judgement, based on the past experience of himself and his colleagues, offers the best prognosis. Since there are few conditions for which treatment is 100% effective any clinician with imagination is always on the lookout for potential improvements in therapy. When a possible new treatment first materializes, the more adventurous and enthusiastic investigators might try it out on a few patients in an uncontrolled trial. That is, the new treatment is studied without any direct comparison with a similar group of patients on more standard therapy. To give the new treatment a reasonable chance of success one might select less seriously ill patients: consequently, regardless of the treatment’s real value such a selected experimental group of patients will appear to do surprisingly well compared with the general routine. Also, one might tend to place greater emphasis on successes, perhaps even exaggerate them a little, and might fail to report some failures on the basis that such patients were clearly ‘too ill’ to benefit from the new treatment. This critical opening paragraph serves to emphasize that uncontrolled trials have the potential to provide a very distorted view of therapy especially in the hands of slipshod, over-enthusiastic or unscrupulous investigators.

Pre-20th century medicine was largely based on such an uncontrolled approach to the promotion of a new therapy but more recent examples may still be found. Advanced cancer is one disease which has frequently experienced extravagant claims for therapeutic effect. For instance, in the United States the drug Laetrile has achieved widespread popular support for treating advanced cancer of all kinds without any formal testing in clinical trials. Ellison et al. (1978) reported an extensive enquiry by the National Cancer Institute to collect well-documented cases of tumour response after Laetrile therapy. Although an estimated 70,000 cancer patients have tried Laetrile only 93 cases were submitted for evaluation of which six were judged to have achieved a response. This examination of such an uncontrolled collection of cases is clearly not good scientific evidence, but did provide some preliminary objective indication that Laetrile is not a ‘cancer cure’ which helped to counterbalance the emotional claims by its advocates. Moertel et al. (1982) have since reported an uncontrolled trial of Laetrile treatment for 178 patients with advanced cancer. The results were not encouraging: the median survival time was 4.8 months and indications of cyanide toxicity occurred in several patients. Since uncontrolled trials are usually over-optimistic, this particular trial offers support to the historical controls and non-randomized concurrent controls. Essentially, the aim is to show that from all such non-randomized studies it is very difficult to obtain a reliable assessment of treatment efficacy. In section 4.4 some of the ethical and practical issues associated with randomization will be discussed. The mechanical details of how randomization is actually prepared and carried out will be explained in chapter 5.
realistic conclusion that ‘Laetrile is a toxic drug that is not effective as a cancer treatment’. The American experience with Laetrile indicates the worst possible situation where a therapy gains wide acceptance by the lay public (though not the medical profession) without any proper evidence of patient benefit.

Interferon, another potential anticancer agent, provides an analogous situation where many clinicians as well as laymen are very enthusiastic about its activity before any properly controlled trials have been performed. Yanchinski (1980) describes the background whereby animal studies, knowledge of its antiviral properties and reports of tumour shrinkage in a few patients have led to the opinion that interferon could be a tremendous breakthrough in the treatment of cancer. Interferon is currently in very short supply such that it can only be tested in a few centres. Studies so far have been uncontrolled and for patients with many different types of advanced cancer. Results are encouraging, but past experience with many other cancer drugs tells one that the early promise shown in uncontrolled studies often fails to be substantiated once properly controlled trials are undertaken. In the case of interferon one must express doubts about how representative the selected cases are and how consistent patient evaluations have been. In Britain, clinical research into interferon had an unfortunate start since the treatment of two Glasgow children was widely publicized. The initial response led to extravagant press claims for ‘cancer cure’ but both children subsequently died. With a drug in such short supply it is scientifically and ethically inexcusable not to undertake randomized controlled trials as soon as possible. More recently, the British Imperial Cancer Research Fund have started a randomized trial of interferon for patients with locally recurrent breast cancer.

With more conventional chemotherapy for the treatment of cancer it has become standard practice to carry out uncontrolled phase II trials of new drugs once phase I trials have established an ‘appropriate’ dose schedule. A separate trial is undertaken for each cancer site in advanced cases and the idea is to see what percentage of patients achieve some objective measure of tumour shrinkage. Only those drugs with an adequate proportion of responders will be studied further in randomized phase III trials. Moertel and Reitemeier (1969) reported the results of 20 different trials of the same treatment (rapid injection of the drug 5-FU) for the same disease (advanced bowel cancer) and their findings illustrate the general difficulty in interpreting such uncontrolled phase II trials. The percentage of responders on these 20 trials ranged from 8% to 85%. Admittedly, these extremes arose from the smaller trials with fewer than 20 patients, but even the six larger trials with between 40 and 150 patients still showed tremendously variable results with response rates ranging from 11% to 55%. Why such incompatible findings for seemingly identical trials? Perhaps the single most important reason is patient selection. Although all patients had advanced colorectal cancer, different investigators will differ as regards the stage of disease progression their patients have reached prior to 5-FU therapy: some will have used 5-FU as a last resort for very advanced patients, perhaps after other drugs have failed, while others will have been more adventurous in using 5-FU as soon as advanced cancer was detected. Other contributory reasons will be variation in the criterion of objective tumour regression and different approaches to the continuance of treatment, especially if drug toxicity occurs. However, whatever the reasons, this example indicates that the response rate for a drug depends very much on who is doing the trial. Nevertheless, most early (phase II) trials to assess the potential of new cancer drugs remain uncontrolled. This undoubtedly means that some ineffective drugs may be over-optimistically reported and also some effective drugs given to very advanced patients may receive inadequate study due to initial poor results. The use of uncontrolled phase II trials for many other conditions (e.g. psychiatric illness) may be totally unjustified if either the definition of the disease or evaluation of patient outcome is less objective than in advanced cancer.

Perhaps this appalling situation illustrated by Moertel and Reitemeier (1969) has improved somewhat in the past decade. Greater attention is now paid to patient factors affecting prognosis, such as prior therapy and performance status, so that better homogeneity and more detailed reporting of patients entered in a trial can be expected. Also, criteria for tumour response and details of the treatment regimens have become more precise. However, there must remain considerable uncertainty as to the value of uncontrolled phase II trials. Williams and Carter (1978), in an article dealing with many aspects of cancer chemotherapy research, discuss several alternative designs for randomized phase II studies. One approach is to assign patients randomly to the new drug or standard drug therapy with the intention of transferring patients to the other therapy if they fail to respond. This has the advantage of encouraging investigators to try a new drug on less advanced patients, and hence giving it a better chance to show its worth, with the reassuring knowledge that all patients will have the opportunity to receive standard therapy if need be. One may adapt this approach by having a majority, say 2/3rds, of patients on the new drug (see section 5.4 for further discussion of such unequal randomization) thus enabling experience in using the new drug to be gained more quickly.

Another approach is to assign patients randomly to one of several new drugs, this being particularly suitable for cancer sites in which there is no effective standard treatment (e.g. lung cancer). Such a trial is randomized, but not controlled. Compared with having a separate uncontrolled trial for every new drug it has the advantage of ensuring a more representative group of patients for each drug, since investigator bias in selecting patients is not drug-specific.

One general finding is that uncontrolled studies are much more likely to lead to enthusiastic recommendation of the treatment as compared with properly controlled trials. For instance, Foulds (1958) reviewed 52 published uncontrolled trials in psychiatry and found that 85% of them reported a therapeutic success whereas in 20 published trials with a control group only 25% reported therapeutic success.

Grace et al. (1966) provide another useful example in a review of 53 studies of portacaval shunt operation for portal hypertension: 32 of these trials were uncontrolled and 75% of them gave a markedly enthusiastic conclusion in their
publication. In contrast, there were only six well-controlled trials, none of which led to marked enthusiasm though three did lead to moderate support for the treatment.

Chalmers and Schroeder (1979) reviewed therapeutic trials published in the New England Journal of Medicine over the previous 25 years. In the years 1953 and 1963 over half the trials were uncontrolled whereas in 1975-1978 the proportion of uncontrolled trials fell to 30%. This encouraging reduction in uncontrolled trials in one major journal is probably reflected in other journals and in clinical research at large, and one hopes that the trend will continue. Chalmers and Schroeder conclude that ‘the studies without controls are not likely to fool anybody’. I very much hope their assertion is true.

Lastly, one unfortunate use of uncontrolled studies by the pharmaceutical industry sometimes occurs after a drug has been approved for marketing. As a promotional exercise a large number of doctors, often general practitioners, are encouraged to use the newly marketed drug in an uncontrolled (phase IV) trial. Such a trial has virtually no scientific merit and is used as a vehicle to get the drug started in routine medical practice. I would not deny that the marketing of new drugs is of tremendous importance to pharmaceutical companies, but it should not be conducted under the disguise of a clinical trial.

4.2 PROBLEMS WITH HISTORICAL CONTROLS

After accepting the need for a control group receiving standard treatment, many researchers are still reluctant to assign patients randomly to new or standard treatment. Such reluctance often stems from the investigator’s desire to enter all future patients on the new treatment because of his wish to gain as much experience of it as possible and his inclination to believe it is better anyway. The most common way of avoiding randomization is to compare retrospectively one’s current patients on the new treatment with previous patients who had received standard treatment, this latter group of patients being commonly known as historical controls.

Such an approach has one major flaw: how can one ensure that the comparison is fair? That is, if the treatment and control groups differ with respect to any feature other than the treatment itself, how can one guarantee that any apparent improvement in patient response is actually due to the new treatment? The potential incompatibility can be divided into two broad areas, patient selection and the experimental environment, each of which may give rise to several sources of bias:

Patient Selection

(1) A historical control group is less likely to have clearly defined criteria for patient inclusion, since such patients on standard treatment were not known to be in the clinical trial when their treatment began.

(2) Since historical controls were recruited earlier and possibly from a different source there may be a change in the type of patient available for selection.

(3) There is the danger that the investigator may be more restrictive, either deliberately or subconsciously, in his choice of patients for a new treatment.

Experimental Environment

(1) One common issue is that the quality of the recorded data for historical controls is inferior, again since such patients were not initially intended to be in the trial. Any clinical trial requires forms designed in advance (see chapter 11) and retrospective extraction of information from routine case notes is unlikely to provide adequate data.

(2) The criteria of response may differ between the two groups of patients. Even if the criteria appear to be the same on paper, those evaluating response on the new treatment may differ in their interpretation of such rules as compared with the earlier evaluators for historical controls.

(3) Ancillary patient care may improve on the new treatment. It is very difficult to ensure that all aspects of managing the patient, other than the treatment under study, remain constant. Patients on experimental therapy in a clinical trial may well have closer observation than would occur for routine standard therapy and if patients are aware and approve of being experimented on this may affect their attitude to disease and hence their subsequent response.

(4) There is a tendency to invalidate more patients on a new treatment than in historical controls. Patients on new therapy who fare badly may be excluded after subsequent enquiry reveals some protocol violation whereas the corresponding exclusion of any historical controls is made difficult since considerable time will have elapsed since they were treated.

The nett result of all these problems is that studies with historical controls tend to exaggerate the value of a new treatment. For instance, Grage and Zelen (1982) report on the development of intra-arterial infusion therapy for the treatment of metastatic colorectal carcinoma to the liver. Several studies with historical controls, involving over 1000 patients, extolled the virtues of this therapy whereas the only randomized trial showed no advantage for intra-arterial infusion chemotherapy as compared with standard systemic chemotherapy. One problem was that the over-optimistic results from the earlier studies made it difficult to recruit patients on to the randomized trial since many clinicians were reluctant to randomize patients to systemic chemotherapy being already (falsely) convinced of its inferiority.

Ingelfinger (1972) refers to this same issue by quoting an example of a trial of hydrocortisone treatment after acute myocardial infarction. Mortality was 14.5% as opposed to 23.2% in a non-randomized control group. The authors believed that this study showed hydrocortisone to be beneficial but went on to say that they hoped their study would lead to large-scale randomized trials of
hydrocortisone. This implies that poorly controlled trials are liable to convince some clinicians that a new treatment is better, but have too great a potential bias to be accepted as good scientific evidence. However, the dilemma is that, once the non-randomized study is completed, there may be great difficulty in undertaking subsequent randomized trials. In this case, one has no means of knowing whether the mortality difference is genuine or not and this places researchers in a quandary over whether it is ethical to undertake future trials with a randomized control group. Hence, trials with historical controls have the tendency to confuse rather than clarify clinical issues and should be avoided even as pilot studies. This has led Chalmers et al. (1972) to advocate that randomization should be introduced in the very earliest clinical trials of a new treatment.

As mentioned in section 2.4, the review of clinical trials for anticoagulant therapy after myocardial infarction by Chalmers et al. (1977) showed that use of historical controls led to the reduction in mortality being greatly exaggerated as compared with randomized trials. This indicates that even with such an objective outcome as death, there is ample scope for bias in non-randomized trials.

Similar exaggeration of treatment benefit is reported by Grace et al. (1966) in their review of trials for portacaval shunt operation mentioned earlier. Out of 15 trials with non-randomized controls ten reported marked enthusiasm for the operation compared with none of the six randomized trials. This indicates that poorly controlled studies are not dissimilar from uncontrolled trials as regards the tendency for over-enthusiastic conclusions.

Thus, there has been increasing scepticism regarding the validity of historical controls and this is reflected in the review by Chalmers and Schroeder (1979) of clinical trials published in the New England Journal of Medicine. Whereas in 1976-1978 only two trials (<1%) had historical controls, this applied to over 10% of trials in earlier years.

Nevertheless, there are researchers who argue in support of historically controlled studies (see Gehan and Freireich, 1974, and Cranberg, 1979). Both articles advocate that historical controls can be of value if sufficient care is exercised in the study’s conduct. Indeed, I would agree with their view that on some occasions the use of historical controls may give an unbiased result. However, when presented with the findings of any particular trial with historical controls I see no way to evaluate whether one has been fortunate enough in this goal. That is, trials with historical controls can never be interpreted with the same degree of confidence as properly executed randomized controlled trials. Byar et al. (1976) and Doll and Peto (1980) are two interesting responses to the above articles, both of which argue in favour of randomized trials.

With the above divergence of medical opinion it seems likely that there will still be some trials with historical controls in the future. Hence it seems relevant to present guidelines as to which of these studies are liable to be the least unacceptable.

Firstly, literature controls, whereby the control group is made up of patients treated elsewhere and previously reported in the medical literature, offer a particularly poor comparison of treatments. They allow ample opportunity for differences in all aspects of patient selection and experimental environment mentioned earlier, so such studies are essentially worthless. Another slightly different problem arises in a review of the literature when the response of several therapies tried in different centres is compared. For instance, Goldsmith and Carter (1974) compared 13 drugs for the treatment of Hodgkin’s disease by tabulating all the available data from uncontrolled phase II trials. Thus, vinblastine had a 68% response rate in 682 patients compared with a 50% response rate in 149 patients on BCNU. However, since both sets of patients are derived from several different studies with differing patient selection and methods of patient evaluation, one cannot really be sure that vinblastine is more active. Such review articles are undoubtedly of interest but need to be interpreted cautiously.

Historical controls obtained from within the same organization might be thought to offer a more reliable comparison but this may not be so if the historical data were not part of a previous trial. For instance, I recall a colleague who wished to compare surgery with more conservative treatment of heel bone fractures. There were 30 new surgical cases and several hundred previous conservatively treated cases. His intention to match each case with a ‘similar’ control can only partly eliminate bias since, although it may largely account for differences in patient selection, the experimental environment including the advantage often associated with being in a trial will remain vastly different. In such instances, the investigator should recognize that he has really conducted an uncontrolled phase II trial which has very limited non-comparative conclusions. Most of the examples earlier in this section derived their historical controls in this way. It might seem the most logical approach, to compare one’s new treatment with one’s own past practice, and indeed it may well lead to a useful learning experience for the investigator concerned. However, it cannot provide a reliable advance in scientific knowledge.

If historical controls are obtained from a previous trial in the same organization one might seem to stand a better chance of reducing the potential bias. One should require that such a previous trial be recent and comparable to the current trial in such features as type of patient and methods of evaluation. However, Pocock (1977b) has shown that there may still be problems. From three cancer cooperative groups in the United States, 19 instances were identified where the same treatment had been used in two consecutive trials. If historical comparisons of this type are without bias, one would not expect any notable difference in survival for the two groups receiving the same treatment. In fact, the 19 changes in death rate ranged from -46% to +24%, and in four instances the difference was statistically significant (each P < 0.02). Thus, even comparisons with one’s previous trial need to be treated with caution.

Byar et al. (1976) illustrate the problem further with an example from a large US multi-centre trial in prostate cancer. This trial showed no survival difference between placebo and estrogen therapy, but if one compared placebo patients in
the first 2½ years with oestrogen patients in the second 2½ years, the latter group had significantly better survival. Although the protocol had not changed, the earlier part of the study had a greater proportion of older patients with poor performance status, such that if the former had been used as a recent historical control for the latter an incorrect inference would have been drawn.

Gehan (1978) has suggested that such historical bias can be overcome by using more complex statistical methods (such as analysis of covariance) to allow for differences in patient characteristics for treatment and control groups. Indeed, Byar et al., go on to state that the above survival difference is removed after adjustment for the prognostic factors age and performance status. Such methods of analysis are described in section 14.1 and also by Armitage and Gehan (1974) in the more general context of how to identify and use prognostic factors in the design and analysis of clinical trials. However, I wish to state several reasons why such retrospective adjustment for trials with historical controls is liable to be unsatisfactory:

(1) Historical data are often of poorer quality so that reporting of prognostic factors may not be consistent.
(2) One may have only a sketchy idea of which patient factors are important and some essential factors may go undetected.
(3) Prognostic factors can only adjust for patient selection, whereas bias due to changes in experimental environment will remain.
(4) The analysis techniques are quite complex and involve certain assumptions, which may not be fulfilled. The methods may be clear to a skilled data analyst but their interpretation might confuse many clinicians.
(5) To propose that poor design can be corrected for by subtle analysis techniques is contrary to good scientific thinking.

Gehan illustrates his approach with a trial of adjuvant chemoimmunotherapy (FAC-BCG) in primary breast cancer. The historical controls who only had a mastectomy were not in a previous trial and had certain major differences from the treatment group: only 22% of controls were treated at M. D. Anderson hospital compared with 47% on FAC-BCG and 55% of controls had a radical mastectomy as compared with 55% on FAC-BCG. Such marked discrepancies indicate that the control group was being handled in a very different manner from the treatment group, and statistical techniques can only partially compensate for this. The apparent superiority of FAC-BCG may well be genuine but we will never know the extent to which poor design led to an exaggeration of treatment benefit. The intent of this trial is very similar to the L-Pam trial described in section 1.4, but I feel the manner of its conduct is a poor substitute for the randomized controlled trial.

Gehan and Freireich (1974) suggest that another means of overcoming historical bias is by matching each new patient with one or more control patients such that they are alike with regard to the major prognostic factors. They go on to describe a trial to evaluate a protected environment for acute leukemia patients reported by Bodey et al. (1971). Each of 33 patients receiving chemotherapy and antibiotics in a protected environment were matched to two control patients. The method of matching was quite complicated and involved some judgement since there were non patient factors the investigators wished to account for. This illustrates the difficulty (indeed impossibility) of achieving a perfect match. The results showed that the protected patients had improved remission and survival and a reduction in infections. However, for reasons (1) to (3) mentioned above I would argue that historical bias may still be present in such a design.

One could argue that in some circumstances the benefit to be derived from a new treatment may be so great that use of historical controls could not seriously mislead. The trouble is that one only knows that a treatment is much superior after a trial has been performed. There are all too many instances where prior to a trial investigators will claim their new treatment as 'the greatest invention since sliced bread' implying that the clinical trial is only a formality, whereas if the trial is properly conducted the eventual findings may show no real benefit. Even if use of historical controls does give the right answer, i.e. a genuinely superior treatment is shown to be better, one still would like to know how much better and the uncertainty of historical bias makes this difficult to assess.

Another argument is that if a disease is rare then one will have difficulty in accumulating enough patients for a randomized controlled trial in which only half the patients receive the new treatment. Here the use of historical controls appears a convenient suboptimal solution leading to quicker results since all new patients receive the experimental therapy. This approach is not totally without foundation, but if sufficient collaborative effort is concentrated on gathering all patients with the rare condition from a large enough population then randomized trials are still feasible. For instance, in the treatment of Wilms' tumour, a rare childhood cancer, a randomized trial has been achieved by a national effort in the United States (see D'Angio et al., 1976).

However, the case for historical controls is stronger for trials with very limited numbers of patients. The larger sampling error in a randomized trial needs to be balanced against the uncertainty of historical comparison and Meier (1975) has considered this concept in a mathematical setting as follows: Suppose historical controls have bias in response represented by a random variable with mean 0 and variance $\sigma^2$ and that sampling variation in response on each treatment is denoted by $\tau^2$. Then if there are $H$ historical controls and $N$ new patients to be entered on trial the choice is between (a) all $N$ patients on the new treatment, or (b) $N/2$ patients on each of the new and standard treatments using randomization. Meier shows that the former, i.e. historical controls, is to be preferred if $\sigma^2 > 2\tau^2/N - \tau^2/H$. In reality one has no simple means of determining $\sigma^2$, but the formula does indicate that the case for historical controls is made stronger as $N$ increases and/or $H$ increases. Such favourable circumstances may exist for small phase II trials when substantial control data are available from previous trials.

This statistical argument has been extended by Pocock (1976) to consider 'unequal' randomization in which more than half the random assignments are
to the new treatment. The optimal solution is then for the number of patients \( R \) on the randomized control group to be \( R = \frac{\kappa}{(1 + H \Delta)} \). The intention would be to include both randomized and historical controls in the eventual analysis of results, though giving more weight to the former. Further comment on the use of unequal randomization to give a greater proportion of patients on a new treatment is given in section 5.4.

### 4.3 PROBLEMS WITH CONCURRENT NON-RANDOMIZED CONTROLS

Even when the investigators have agreed to a prospective clinical trial in which future patients are to be assigned to the various treatments, there may still be some reservations about whether such assignment should be based on a random mechanism. Instead, it may be decided to use some predetermined systematic method, or worse still some degree of judgement by investigator and/or patient may be adopted. This section is concerned with the problems that can arise from using such concurrent non-randomized controls.

#### Systematic Assignment

The most common methods used here are to assign patients according to the date of birth (e.g. odd/even day of birth = new/standard treatment) or date of presentation (e.g. odd/even days = new/standard treatment) or to use alternate assignment (e.g. odd/even patients = new/standard treatment). The main problem with all of these methods is that the investigator can easily know in advance which treatment a patient would receive if he entered the trial and this prior knowledge may affect the investigator’s decision regarding entry or not.

For instance, Wright et al. (1948) report on a trial of anticoagulant therapy for myocardial infarction whereby patients admitted on odd days of the month received anticoagulant and patients admitted on even days did not. There were 589 treated and 442 control patients, a sizeable imbalance indicating a preference towards admitting patients onto anticoagulants. This finding brings into question the comparability of the treatment and control groups and hence the validity of the results.

Similarly, Grage (1981) reports a trial of preoperative radiotherapy for rectal cancer begun in 1957 in which patients were assigned according to birth date: 192 patients received preoperative radiation compared with 267 treated by operative resection alone, again an imbalance which casts doubt on the trial’s validity.

In the case of alternate assignment it is somewhat more difficult to detect bias, since although the investigator’s prior knowledge of the next treatment may affect patient selection the equality of treatment numbers will be preserved. However, one may find some lack of comparability in the characteristics of the treatment groups. For instance, Ehrenkranz et al. (1978) evaluated vitamin E for neonates with bronchopulmonary dysplasia by alternate assignment of 40 such infants to vitamin E and control groups. The vitamin E group had a higher mean 1-minute Apgar score, which raises the possibility that there might have been some selection bias so that the trial’s findings in favour of vitamin E cannot be interpreted with quite the same confidence as if the trial was randomized. Of course, even if random assignment is used one can still get chance differences in treatment groups, but provided randomization is arranged so that investigators do not know which treatment is coming next no selection bias is possible (see chapter 5).

Thus, there would seem no real justification for such systematic assignment methods since they do contain a potential bias and can be replaced quite simply by randomization.

Another potentially more serious problem arises if a trial is conducted so that the treatment depends on the clinician, whereby some clinicians (or hospitals) give one treatment while other clinicians (or hospitals) give another. This approach has much the same deficiencies as historical controls since both patient selection and the experimental environment may differ considerably between treatments.

Cockburn et al. (1980) conducted a trial of vitamin D supplement versus placebo in pregnant women to see if vitamin D could reduce neonatal hypocalcaemia. Mothers assigned to one hospital ward received vitamin D while mothers admitted to another ward did not. Patients in the two groups were comparable for social class, parity and maternal age so that patient selection appeared no problem. However, the two wards were under the care of different consultants and this raises the possibility that the vitamin D group could have differed from the controls in some other aspects of medical care. The results showed marked benefits in the vitamin D group, but having such non-randomized controls leaves some doubt. In a subsequent larger trial with a higher dose of vitamin D the investigators have implemented randomized assignment to overcome such qualms.

The wish to compare different treatments given in different hospitals can arise if each hospital is committed to a certain fixed approach. For instance, clinical trials for the evaluation of radiotherapy for cancer in the United States have at times been difficult to get off the ground since many cancer centres are unwilling to deviate from their standard treatment. Schoenfeld and Gelber (1979) mention an unusual way round this problem whereby, in a trial with more than two treatment options, each centre could opt out of certain treatment(s) they disapprove of, and have each of their patients randomized to the remaining options. Of course, it would be better if all centres could agree on the treatments to be compared, but perhaps a randomized trial which allows options is better than a non-randomized trial or no trial at all.

#### Judgement Assignment

If the investigator and/or the patient is allowed to exercise his judgement in assigning one of several treatment options it is evident that this could introduce
considerable bias: for instance, the investigator may favour one particular treatment for his more serious cases which is liable to make this treatment appear worse regardless of its actual merit. Hence, such use of judgement is generally regarded as obviously unacceptable in clinical trials and one will see few explicit examples of its use in the medical literature.

However, one instance reported by Smithells et al. (1980) is a trial of vitamin supplementation for prevention of neural tube defects (NTD) given to high-risk women planning a further pregnancy. Here the untreated control groups included some women who had declined to take vitamin supplements (i.e. patients were effectively allowed to choose whether they were in the treatment or control group) as well as women who were already pregnant. Lack of randomization in this trial has made it impossible to decipher whether the reduced number of NTD infants after vitamin supplementation is really due to the vitamins themselves or due to bias in patient selection. The ensuing controversy has hampered plans by the Medical Research Council to run a randomized controlled trial which could properly resolve this issue.

It should also be noted that the use of judgement in treatment assignment may still be present even when not explicitly mentioned. For instance, if a report of a clinical trial merely provides a comparison of two or more treatments with no indication as to how patients were assigned one should not automatically assume that judgement played no part.

Another problem is where the investigators interfere with a randomized trial. ‘Student’ (1931) describes one such classic example in the Lanarkshire milk experiment. In 1930, 10 000 children received 3/4 pint of milk a day at school while another 10 000 in the same schools did not, the objective being to see if such milk supplement led to increased height and weight. However, trouble arose in the trial’s design as ‘Student’ explains:

The teachers selected the two classes of pupils, those getting milk and those acting as controls in two different ways. In certain cases they were selected by ballot and in others on an alphabetical system. So far so good, but after invoking the goddess of chance they unfortunately wavered in their adherence. In any particular school where there was any group to which these methods had given an undue proportion of well fed or ill nourished children, others were substituted to obtain a more level selection. This is just the sort of afterthought that is apt to spoil the best laid plans. In this case it was a fatal mistake for in consequence the controls were definitely superior both in weight and height by an amount equivalent to about 3 months’ growth in weight and 4 months’ growth in height. It would seem probable that the teachers swayed by the very human feeling that the poorer children need the milk... must have unconsciously made too large a substitution of the ill-nourished among the ‘feeders’.

Those children receiving milk tended to gain more height and weight, but the initial differences cast doubt on the extent to which this could be attributed to the milk itself.

In recent years the use of data banks on computer containing information on all previous patients in a given institution has been advocated by some enthusiasts as an exciting development in clinical research. For instance, Starmer et al. (1974) describe the use of data banks in the management of chronic illness. It has been proposed that such data banks could be used in the evaluation of different treatments and might be a substitute for randomized clinical trials. However, Byar (1980) provides a firm rebuttal of such an idea. Basically, such retrospective comparisons of treatment from a data bank arise after several clinicians have used their judgement in deciding which treatment their patients should receive. Also, the lack of any precise protocol means that treatments, types of patients and methods of evaluation cannot conform to any consistent definition. Thus, although they may provide a useful insight into the general pattern of patient management and prognosis as experienced in one institution data banks provide very poor quality information for treatment comparison, perhaps even worse than the historical controls I criticized in section 4.2.

4.4 IS RANDOMIZATION FEASIBLE?

So far I have shown that various alternatives to randomization are liable to produce seriously biased and often overoptimistic results regarding a new therapy. Hence on purely scientific grounds it is easy to deduce that the use of a randomized control group is to be preferred in all situations. Furthermore, in more practical terms randomized controlled trials are an efficient method for determining the optimal therapy for future patients. However, clinical trials are not solely to do with the advancement of scientific knowledge and one needs to take into account the actual circumstances regarding eligible patients before automatically proceeding with a randomized trial. In particular, one must consider whether it is ethical to randomize patients and also whether there are sufficient investigators (and hence patients) willing to participate in such randomization.

As regards the ethics of randomization, Hill (1963) provides a carefully reasoned argument. He begins by considering the first randomized trial, to evaluate streptomycin treatment for pulmonary tuberculosis, already mentioned in chapter 2. Streptomycin was in short supply at the time so that one could not have given it to all patients even if one wanted to. Since efficacy had not been previously established, Hill argues it would have been unethical not to seize the opportunity to conduct a randomized controlled trial. Such a situation of therapy in short supply (interferon mentioned in section 4.1 is another example) makes it particularly easy to randomize since in addition to scientific validity one is also exercising ‘fairness’ in giving each patient an equal chance of receiving the rare treatment.

However, the ethics of randomization require a more subtle argument if a new therapy is in plentiful supply and could, if one wished, be given to every new patient. First, one assumes that the wish to conduct a clinical trial is based on the idea that the new treatment has a reasonable chance of being a genuine improvement. Indeed, one must expect that some clinicians will already be inclined to believe that the new treatment is better. However, opinion should
not necessarily preclude an investigator from entering patients on a randomized trial. It is important to draw a distinction between such subjective personal belief and objective scientific knowledge regarding efficacy. For instance, Gilbert et al. (1977) reviewed 46 randomized controlled trials in surgery and anaesthesia and found that in only half of such trials was the therapeutic innovation found to be preferable. The motivation behind each of these trials was a belief that the innovation was liable to be better, but it turns out that there is a substantial probability that such prior expectations will not be fulfilled.

Hence, before agreeing to enter patients in a randomized trial each investigator must come to terms with his personal judgement of the treatments involved. It is inevitable that he will have some preferences regarding treatment, but past experience (as in the above example) shows that randomized trials have a habit of often producing scientific evidence which contradicts such prior belief. Of course, if a clinician feels very strongly that one treatment in a randomized trial is unacceptable then he should not participate. However, it is for each clinician to decide whether he has the right to take such a dogmatic stance or needs to have his beliefs checked empirically by a randomized trial. Further consideration of ethical problems in clinical trials is given in chapter 7.

One further problem is in deciding when is the opportune moment in the development of a new therapy to start a randomized trial and let us consider this issue as regards surgical trials. Chalmers (1972) argues that randomization is introduced infrequently and too late to evaluate new operations. For instance, he refers to 152 trials of operative therapy for coronary artery disease: only two trials were randomized and both found internal mammary artery ligation of no value. He argues that ‘the only way to avoid the distorting influences of uncontrolled trials is to begin randomization with the first patient’. However, Bonchek (1979) and Van der Linden (1980) discuss some of the difficulties associated with randomized surgical trials. In particular, the skill of the surgeon is likely to have an impact on the patient’s prognosis, more so than the clinician’s impact in a drug trial. Thus, in comparing a new surgical procedure against non-surgery the former will be going through phases of development whereby refinements in technique will often mean that later results may surpass the achievements of the first experimental operations. Also, caution is needed in generalizing from the achievements of the most skilled and experienced surgeons in an experiment to the lesser expectations of routine surgical practice.

If a randomized trial is performed after a treatment has become standard practice then its results are likely to provoke controversy. The Danish Obesity Project (1979) compared a widely accepted surgical procedure (jejunooileal bypass) with medical treatment of morbid obesity and found the former to produce greater weight loss and improved quality of life, though with complications of surgery common and sometimes severe. Since the trial confirmed a suspected benefit of surgery, it might be considered unnecessary and unethical. But how is one to know such benefit if a randomized trial is never performed? Evidently, controversy would be avoided if the development of new surgery involved randomized trials at an earlier stage.

The comparison of alternative surgical procedures raises the problem that surgeons may be more experienced in one operation than the other, and such difference in pretrial routine may affect the results of a trial. For example, Van der Linden (1980) refers to two trials, one Swedish and one Finnish, which studied early versus delayed operation for acute cholecystitis and which came to completely opposite conclusions. Again, I do not think such contradiction is an argument against doing randomized surgical trials but more an indication that the relevance of their findings must be assessed relative to each hospital’s circumstances. Grage and Zelen (1982) point out that randomization may be especially difficult if the treatment modalities are radically different. For instance, in the management of soft tissue sarcoma one would like to compare local excision plus radiotherapy with a more radical excision or amputation. However, it would be extremely difficult to conduct a trial in which whether the patient loses an arm or leg depends on random assignment. Here one may have to resort to some form of non-randomized comparison, though use of a randomized consent design (see section 7.2) is a possibility.

The purpose of this chapter has been to emphasize that in general randomized controlled trials are an essential tool for testing the efficacy of therapeutic innovations. The proper use of randomization guarantees that there is no bias in the selection of patients for the different treatments and also helps considerably to reduce the risk of differences in experimental environment. Randomized allocation is not difficult to implement and enables trial conclusions to be more believable than other forms of treatment allocation.

However, the acceptance of randomization remains only a starting point in the proper execution of a trial. In particular, if the randomization is not performed correctly then there is every danger that the trial might be just as biased as the non-randomized trials mentioned earlier. Hence, the next chapter describes various methods of implementing random treatment assignment and discusses some of the pitfalls to be avoided.