
RANDOMISED CONTROLLED CLINICAL TRIALS

CHRISTOPHER J. BULPITT

Senior Lecturer in Epidemiology,
Department of Medical Statistics and Epidemiology,
London School of Hygiene and Tropical Medicine,
Keppel Street,
London WC1
Honorary Senior Lecturer in Clinical Pharmacology,
Royal Postgraduate Medical School, London
and Honorary Consultant Physician,
Hammersmith Hospital, London



MARTINUS NIJHOFF PUBLISHERS
THE HAGUE/BOSTON/LONDON

Copyright 1983 © by Martinus Nijhoff Publishers, The Hague

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, mechanical, photocopying, recording, or otherwise, without written permission of the publisher, Martinus Nijhoff Publishers, P.O. Box 566, 2501 CN The Hague, The Netherlands.

Distributors:

for the United States and Canada

Kluwer Boston, Inc.
190 Old Derby Street
Hingham, MA 02043

for all other Countries

Kluwer Academic Publishers Group
Distribution Center
P.O. Box 322
3300 AH Dordrecht
The Netherlands

Bulpitt, Christopher J.
Randomised controlled trials.

(Developments in biostatistics and epidemiology; v. 1)

1. Medical research—Methodology. 2. Human experimentation in medicine. I. Title. II. Series. [DNLM: 1. Clinical trials. 2. Research design. 3. Biometry. 4. Epidemiology. W1 DE997VL v. 1 / W 20.5 B939r] R850.B83 1983 615.5'0724 82-14544
ISBN 90-247-2749-9

CONTENTS

Preface vii

1. Introduction 1 G ✓
2. The history of controlled trials 5 G ✓
3. Ethical considerations 12 G ✓
4. The objectives of a randomised controlled trial 28 G
5. Validity of the results 35 G
6. Recruitment of subjects 39 G
7. How to ensure that the control and treated patients are similar in all important respects 44 G
8. How to ensure that the results are free of bias 56 G
9. The variability of results 80
10. How many subjects are required for a trial? 96 G
11. Different trial designs 118
12. Writing the protocol 136
13. Information to be collected during a trial 144
14. The conduct of the trial 157
15. Analysis of the trial results 179
16. The evaluation of subjective well-being 194 G

- 17. Early trials on new drugs 209 6
- 18. The detection of adverse drug reactions 214 6
- 19. Failure to accept the results of randomised controlled trials 221
- 20. The advantages and disadvantages of randomised controlled trials 239 6
- 21. References 244

PREFACE



Bradford Hill has defined a clinical trial as "A carefully and ethically designed experiment with the aim of answering some precisely framed question" [1]. This definition specifies a careful design and requires the provision of adequate controls. Random allocation of treatments to subjects is important to ensure that the treated and control groups are similar. Therefore this book is entitled *Randomised Controlled Clinical Trials*. We can define a randomised controlled trial by rewriting Bradford Hill's definition as follows, "A carefully and ethically designed experiment which includes the provision of adequate and appropriate controls by a process of randomisation, so that precisely framed questions can be answered."

I am a firm advocate of Randomised Controlled Clinical Trials but intend to give a balanced view of the advantages and disadvantages of these ethical experiments. This book is directed primarily at the medical research worker, although certain chapters may find a wider application.

When discussing a randomised controlled trial, it is neither practicable nor desirable to divorce theory from practice, however the first ten chapters concentrate mainly on theory, and the remainder focus on practice. The segment on trial design is followed by sections on writing the protocol, designing the forms, conducting the trial, and analysing the results. This book is meant to serve both as a reference manual and a practical guide to the design and performance of a trial.

1. INTRODUCTION

1.1 DEFINITION

A randomised controlled trial was defined in the preface as "A carefully and ethically designed experiment which includes the provision of adequate and appropriate controls by a process of randomisation, so that precisely framed questions can be answered." In medical research, treatment is allocated to subjects or certain periods of time by a random (chance) procedure.

1.2 WHY PERFORM A RANDOMISED CONTROLLED TRIAL?

The primary objective in writing this book was to demonstrate the importance of performing randomised controlled trials and the second was to help in the design and performance of such trials. I do not expect to convince the reader of the necessity for randomised controlled trials in a brief introduction as an entire half of the book is intended for this; but a preliminary discussion may be appropriate. Any reader already convinced of the necessity for randomised clinical trials should proceed to section 1.3.

In the current medical literature, opinion is often held in high esteem and randomised controlled trials constitute only a small proportion of research reports. Articles consisting of observations without randomised comparison groups can be valuable and often generate hypotheses, some of which are subsequently tested by randomised controlled trials, but the preponderance of observational studies over controlled experimentation is surprising. I leave the

reader to imagine the number of controlled trials to be found in literature on the social sciences [my italics].

Cochrane [2] observed that not only were randomised controlled trials neglected in other fields but that in medicine these trials are carried out in developed, capitalist, predominantly Protestant countries. He understood why underdevelopment mitigated against such trials and could only speculate why Communist or Roman Catholic countries should inhibit the performance of randomised controlled trials.

Every time a treatment is prescribed for a patient, whether pharmaceutical agent, operation, diet, psychological counselling, physiotherapy, or other health care strategy, the medical practitioner is conducting a trial of treatment in that patient. Similarly, when an administrator organises health care for a community, for example screening, immunisation, or better housing, an experiment or trial is performed. However, we wish to know whether or not the experiment works and it is not sufficient to observe that the health of the patient or community improves as such improvement may have nothing to do with the experiment. Patients may get better without treatment and the health of a community may improve without screening for diseases. We therefore need controls who do not receive the intervention. These controls provide the baseline against which treatment or intervention can be assessed: hence the term *controlled trials*. The control group should be the same as the intervention group in all respects apart from the intervention procedure. Chapter 7 is devoted to demonstrating that controls must be observed concurrently with the intervention group and that those treated and those serving as controls must be determined by chance alone. The allocation by chance is known as randomisation, leading to the term *randomised controlled trial* or RCT. The allocation can be performed by tossing a coin on each occasion or more usually by the use of random number tables (chapter 7).

1.2.1 Trials of new therapy

Randomised controlled trials are necessary to prove the effectiveness of new health care strategies or treatment and to prevent the introduction of new but useless treatments.

1.2.2 Trials of accepted treatment

Many currently accepted treatments require proof of their effectiveness and trials are still necessary. Such treatments fall into three groups: first, a therapy may have been introduced prior to the advent of clinical trials. Cochrane [2] suggests that psychotherapy, physiotherapy, and surgery for carcinoma of the bronchus can be included in this group. Second, there may be experimental proof of biochemical, psychological, or other effect, but no evidence that the treatment does more good than harm as long-term therapy. Examples are provided by anticoagulant therapy for the secondary prevention of myocardial infarction (section 19.4) and oral hypoglycaemic drugs for maturity-onset

diabetes mellitus (section 19.6). Third, a treatment may have been subjected to randomised controlled trials, but the results are equivocal. Cochrane suggests tonsillectomy as an example of this group.

1.2.3 The place of randomised controlled trials

The place of randomised controlled trials must not be exaggerated. There was no necessity for a trial of streptomycin in tuberculous meningitis; one survival in an otherwise uniformly fatal condition was very conclusive! However, randomised controlled trials of streptomycin were most useful in pulmonary tuberculosis [3]. These trials could have delayed the introduction of active treatment, but a shortage of streptomycin proved to be the limiting factor. Also, any delay may be of value when the active treatment proves to be toxic. The advantages and disadvantages of clinical trials are discussed further in chapter 20. Trials may produce erroneous results but these occur far less frequently than in uncontrolled observational studies. Moreover, the quality of medical care given to patients in clinical trials is much higher than for the usual processes of medical care [4].

1.3 DESCRIPTION OF THIS BOOK

This book is intended as a reference manual for research workers involved in randomised controlled trials and is aimed at the field of medical research. It is hoped that the contents of the book are also relevant to the needs of dental, veterinary, and social science research workers. Also many aspects of trial design have been employed initially in agricultural experiments.

After this introduction, the book considers historical aspects of the subject and the ethical aspects of trial design. We have first to agree the ground rules for trials in patients and in normal paid volunteers. What risks are allowable for the former, if any? It has been suggested that it would be unethical to perform a trial involving the transmission of infectious hepatitis in man, yet such a trial was carried out.

Following the discussion on ethics, subsequent chapters will closely define the trial objectives, validity, recruitment, randomisation, freeing observations from bias, the variability of results, and the numbers required for the trial. The remainder of the book is concerned with practical matters: specific trial designs, writing the protocol, designing the documents, conducting the trial, analysing the results, trials to measure the quality of life, trials on new drugs, the detection of adverse reactions, why results are not accepted, and the advantages and disadvantages of randomised controlled trials.

It is hoped that readers who are not involved in the performance of RCTs will find themselves better able to assess the results of such trials. More importantly, promotional studies, masquerading as important randomised trials, may thus be given the scant attention they deserve. We will attempt to identify the necessary qualities of a satisfactory trial so that we may more readily assess any results. Initially, we will ask if the trial incorporates satisfactory controls.

A trial that is not controlled is often called a study and the results most favourable to a particular treatment usually derive from an open evaluation. An example is given by an article sent to general practitioners by a pharmaceutical company entitled "—— in hypertension. General Practice Study. Preliminary report on 717 patients treated originally with methyldopa." The dose of methyldopa in these patients was reduced but not stopped and —— was started. The article stated "Whatever the reasons, there is good evidence that in the majority of this group, control of the blood pressure improved when the dose of methyldopa was reduced and —— was substituted. After —— was introduced there was a reduction in unwanted side effects, and four out of five patients reported subjective improvement." This was an open evaluation with a predictable result. Give a new drug with enthusiasm and the patient will feel better. This study made nearly 600 patients feel better, but some or all of the improvement may have been due to the attention of the general practitioner rather than the new drug. An article in the *Sunday Times* of January 29, 1978 attacked these marketing trials and put it more clearly. "If the doctor believes the new drug may help, he will almost certainly tell the patient, so the patient will be inclined to prefer it to his previous drug, which may be just as good. Few of these trials make a proper scientific comparison of the new drug with other drugs or with a dummy tablet." The newspaper article went on to consider the profits to be made by pharmaceutical companies who introduce new drugs to general practitioners in this manner if the drug continues to be prescribed after the end of the trial. Not every one agrees that making a profit is undesirable, especially in view of the therapeutic advances made by the pharmaceutical industry. However, there must be adequate proof of benefit and a randomised controlled trial is the method of choice.

2. THE HISTORY OF CONTROLLED TRIALS

2.1 THE EARLIEST TRIALS

The first well documented randomised controlled trial of medical treatment may have been that organised by the Medical Research Council and reported in 1948 [3]. However, Rose and Armitage have described a possible RCT dating from 1662 [5] and R.A. Fisher introduced randomised trials into agricultural research in 1920.

Nonrandomised trials date back many years. L'Etang [6] considered that the story of Daniel contained a report of a clinical trial. Nebuchadnezzar II organised the trial by giving youths of royal blood, including Daniel, a rigid diet of meat and wine for three years. The trial was supervised by a eunuch [monitor]. Nebuchadnezzar's trial was not controlled but Daniel "persuaded the monitor to give him and three others a diet of pulse and water for 10 days." L'Etang reported that these four were "fairer in countenance and fatter in body than the other subjects who were given meat and wine" and concluded that "Daniel had ruined the trial . . . and the trial had become uncontrolled." Daniel had not ruined the trial but had performed one of the first *controlled* trials: a within-subject cross-over study.

Bull has reviewed the history of clinical trials and I am indebted to him for much material in this chapter [7]. Bull cited a second unintentional trial by Ambroise Paré. In 1537 Paré was responsible for the treatment of numerous wounded and ran out of boiling oil used for cauterising the wounds. He was "constrained to apply in its place a digestive made of yolks of eggs, oil of roses

and turpentine." The following day he was surprised to find that those receiving the new medicant "feeling but little pain, their wounds neither swollen nor inflamed . . ." Those who received boiling oil "were feverish with much pain and swelling about their wounds." Paré concluded that the digestive was superior to burning oil but perhaps we would now suggest a longer period of observation would be appropriate in view of the likelihood of subsequent sepsis.

2.2 SCURVY

Bull also reported an unintentional trial from 1600 on an expedition to India by the East India Company. Only one of four ships had lemon juice provided. The ship in question was almost free from scurvy yet the condition was rampant on the other three ships. The company provided lemon juice on all its ships thereafter but presumably this preventive treatment was not fully accepted until 150 years later, when James Lind performed a controlled trial. Bradford Hill quotes James Lind in his book *Statistical Methods in Clinical and Preventive Medicine* [8].

On the 20th May, 1747, I took twelve patients in the scurvy, on board the Salisbury at sea. Their cases were as similar as I could have them. They all in general had putrid gums, the spots and lassitude, with weakness of their knees. They lay together in one place, being a proper apartment for the sick in the fore-hold; and had one diet common to all, viz. water-gruel sweetened with sugar in the morning, fresh mutton-broth often times for dinner; at other times puddings, boiled biscuit with sugar etc. and for supper, barley and raisins, rice and currents, sago and wine, or the like. Two of these were ordered each a quart of cyder a day. Two others took twenty-five gutts of elixir vitriol three times a day, upon an empty stomach; using a gargle strongly acidulated with it for their mouths. Two others took spoonfuls of vinegar three times a day, upon an empty stomach; having their gruels and their other food well acidulated with it, as also the gargle for their mouths. Two of the worst patients, with the tendons in the ham rigid (a symptom none of the rest had) were put under a course of sea-water. Of this they drank half a pint every day, and sometimes more or less as it operated by way of gentle physic. Two others had each two oranges and one lemon given them every day. These they ate with greediness, at different times, upon an empty stomach. They continued but six days under this course, having consumed the quantity that could be spared. The two remaining patients, took the bigness of a nutmeg three times a day of an electuary recommended by a hospital-surgeon, made of garlic, mustard-seed, rad. raphan, balsam of Peru, and gum myrrh; using for common drink barley-water well acidulated with tamarinds; by a decoction of which, with the addition of cremor tartor, they were greatly purged three or four times during the course.

The consequence was, that the most sudden and visible good effects were perceived from the use of the oranges and lemons, one of those who had taken them, being at the end of six days fit for duty. The spots were not indeed at that time quite off his body, nor his gums sound, but without any other medicine, than a gargle of elixir vitriol, he became quite healthy before we came into Plymouth, which was on the 16th June. The other was the best recovered of any in his condition; and being now deemed pretty well, was appointed nurse to the rest of the sick.

James Lind showed the superiority of citrus fruits in the treatment of scurvy. Interestingly, he commits one fundamental error: he appears to have given two of the worst patients a particular treatment (sea water). Perhaps sea water was his favourite treatment. If this was the case, and we assume that more severely affected patients are less easy to cure, the provision of a favourite treatment for these patients will mitigate against demonstrating a benefit. Random allocation to treatment groups prevents this difficulty; this advance did not occur until 1948.

2.3 VACCINATION AGAINST SMALLPOX

In the early eighteenth century inoculation against smallpox with live virus was introduced from Constantinople by Maitland and Lady Mary Wortley Montagu. (They, of course, knew nothing about viruses and reported the inoculation of "smallpox matter".) They arranged for six convicts to be inoculated; all survived and later one was exposed to smallpox and found to be immune [9]. This trial did not reveal that the use of live virus could frequently lead to death and that those inoculated could be infectious, thereby leading to an increase of smallpox in the community.

A similar trial, but using cowpox matter, was next performed by a farmer, Benjamin Jesty in 1774 [10]. Country folk reported that if they had cowpox they would not get smallpox and that cowpox was a mild disease. Farmer Jesty vaccinated himself, his wife, and two children with cowpox material using a stocking needle. Apparently the children were later inoculated with smallpox and were unaffected. Mr. Jesty proceeded to vaccinate his milkmaids but his neighbours considered that such a "bestial" manifestation of smallpox should not be given to man! Benjamin Jesty countered that if we were prepared to eat beef and drink milk, we could be vaccinated with cowpox.

Jesty had performed a remarkable trial and 20 years later Edward Jenner, not knowing of farmer Jesty, was considering the problem and remembered the words of his teacher, the famous surgeon, John Hunter: "Don't think. Do an experiment." Jenner proceeded to do a similar trial which he wrote up in 1798 in *An Inquiry into the Causes and Effects of the Variolae Vaccinae* [11]. He described 23 patients with cowpox who were resistant to smallpox inoculation and assumed that all persons who had neither contracted cowpox nor smallpox would react positively to inoculation with smallpox matter. The trial was not controlled in the strict sense of the word, and controls were necessary owing to the possibility of either natural immunity or previously acquired immunity. Such persons could be vaccinated with cowpox and be immune to smallpox without cause and effect.

Jenner was partly aware of this problem and stated, "To convince myself that the variolus matter made use of was in a perfect state, I at the same time inoculated a patient with some of it who had never gone through the cowpox, and it produced the smallpox in the usual manner." A further problem arose from interpreting resistance to the disease. We now know that inoculating with smallpox would be unethical as a control procedure. Pearson [12] was less

interested in the resistance of his subjects than the state of his variolus matter and made further controlled observations, also in 1798. He observed three patients who had had cowpox and two who had not. Only the three who had had cowpox were immune to inoculation. Interestingly and commendably, Pearson recommended "well-directed observation in a thousand cases of inoculated cowpox." Waterhouse performed a similar controlled trial published two years later [13]. He employed the same number of controls (two) but had 12 in his treatment group.

2.4 THE USE OF PLACEBOS IN THE NINETEENTH CENTURY

When there is doubt about the effectiveness of a particular treatment and when alternate effective treatment is either not available or not required in the short-term, the modern controlled trial may employ a period of placebo medication. In 1801 Haygarth was one of the first to employ placebo treatment [14]. He used dummy appliances to investigate the effects of Perkin's tractors, whose metal rods were supposed to cure by an electrical influence. Haygarth studied five patients; four were helped by the use of wooden imitation tractors. Not only was the trial possibly the first to use placebo medication, but Haygarth quoted Lind, thus "an important lesson in physic is here to be learnt, viz. the wonderful and powerful influence of the passions of the mind upon the state and disorders of the body. This is too often overlooked in the cure of disease . . ."

In 1865 Sutton published a trial of mint water in 20 patients with rheumatic fever [15]. He used mint water, not as an active but as a placebo treatment. On observing a marked tendency to spontaneous cure he remarked, "the best treatment for rheumatic fever has still to be determined." Sutton had been examining the reports of Dr. Gull and he also reported, "No selection was made, but that Dr. Gull treated the cases which happened to be admitted into his wards on the same plan; and we would further beg to say that these reports were not kept for any special object, nor are they as complete as they might be; yet the facts stated, may be fully relied upon, and so far answer our purpose."

These interesting admissions make us suspect that the study was not so well conceived and prospective in design as it first appeared. However, the honesty of Dr. Sutton led to further qualifications that we now agree would be unlikely to influence the course of the disease. ". . . these cases cannot be considered to have been treated solely on the expectant plan, for an occasional dose of Dover's powder or half a grain of opium, right or wrong, and two or three ounces of brandy a day, are remedies that might be fairly expected to exercise some, although, perhaps, little influence over the course of the disease." He wisely stated, "Therefore, cases treated, as the following cases have been, by such simple means that we might almost consider them to be unassisted by any remedy, are invested with no little interest . . . the results . . . will probably warrant us concluding that we ought not to be too hasty in considering the apparent sudden and favourable change in the symptoms due to any medicine administered."

2.5 SCIENTIFIC BASIS FOR CONTROLLED TRIALS

Although the advent of the randomised controlled trial had to wait until the twentieth century, much of the related scientific thinking was published in the nineteenth century. Laplace thought that probability theory ought to be extended to help explain the results observed in medical practice [16], and P.C.A. Louis advocated numeracy in assessing results [17]. Louis stated, "As to different methods of treatment, if it is possible for us to assure ourselves of the superiority of one another among them . . . it is doubtless to be done by enquiring if . . . a greater number of individuals have been cured by one means than another. Here again it is necessary to count." He went on to consider the necessity for controls "in order that the calculation may lead to useful or true results . . . we ought to know the natural progress of the disease." He also appeared concerned about noncompliance: "we ought to know . . . whether the subjects have not committed errors of regimen." Amusingly, he thought his numerical method offered "real difficulties in its execution . . . this method requires much more labour and time than the most distinguished members of our profession can dedicate to it. But what signifies this reproach, except that the research of truth requires much labour, and is beset with difficulty."

Louis used his numerical method in investigating the effect of venesection in 78 cases of pneumonia [18]. Some patients were bled and others were not. Louis not only examined mortality but also symptoms and signs and concluded that bleeding made no difference in outcome. This result was not in keeping with the medical practice at this time, and, not unexpectedly, caused an uproar. However, his findings came to be accepted, a triumph for the clinical trial.

2.6 THE PROVISION OF HISTORICAL CONTROL GROUPS

By the middle of the nineteenth century, rigorous methods of observation had been defined, the necessity for controls realised, and even the statistical theory of probability could have been used in the analysis of results. However, the selection of controls still led to biased results. Elisha Bartlett [19] described the essential requirements for control and treated patients: they should

1. have equal disturbing factors of location, social class, and the like
2. be susceptible of a clear and positive diagnosis
3. not be selected
4. be subjected to a clearly defined method of treatment.

The provision of a more appropriate control group came to be recognised as important and trials started to employ carefully followed historical controls. In 1870 Lister [20] compared the mortality of 35 historical controls with 40 patients treated with antiseptics. Forty-three percent of the controls died but only 15 percent of the treated group. Bull [7] first pointed out that Lister was cautious about drawing conclusions and, second, that more appropriate con-

trols "might have prevented the bitter and profitless controversy which raged for many years." Lister stated, "These numbers are, no doubt, too small for a satisfactory statistical comparison . . ." Bull pointed out that, "The chi-squared test shows them to be highly significant"; perhaps the controversy would have raged less if Lister were a more effective lecturer and a more dogmatic writer. However, historical controls are not appropriate for randomised controlled trials. Pocock [21] has listed 19 instances of the same intervention being used twice in consecutive groups of patients in the same institution. In four of these instances mortality was significantly different between the groups (see also section 8.2.3).

2.7 THE PROVISION OF CONCURRENT CONTROL GROUPS

The lack of acceptance for Lister's trial can be contrasted with that of Pasteur's vaccine for the prophylaxis of anthrax in animals. Pasteur used 60 sheep in the experiment; 25 were inoculated and then infected and 25 were not inoculated but were infected. An additional ten sheep were neither inoculated nor infected. Chance allocation appears to have been employed to some extent in this trial as critical observers suggested the order in which pairs of inoculated and control animals should be infected [22]. All the animals who had been inoculated survived; the 25 controls died. The results of this trial were immediately accepted.

Another early controlled trial was performed by Fibiger. In 1898 he reported a trial of anti-diphtheria serum in alternate patients [23]. He studied 488 patients and showed a reduction in mortality in the patients treated with serum. He also recorded the fact that the diphtheritic membrane disappeared quicker in the treated cases.

In 1945 a trial of penicillin in the treatment of wounds was attempted in the 21 Army Group [24]. The control group was to be those who were given any alternative treatment. Unfortunately, the surgeons were unwilling to withhold penicillin in the presence of serious wounds and the group treated with penicillin were more seriously affected. Despite this bias, the wounds healed quicker in the penicillin-treated group.

2.8 THE DELIBERATE USE OF RANDOMISATION TO PRODUCE SIMILAR TREATMENT AND CONTROL GROUPS

The first modern trial to deliberately employ randomisation may have been the Medical Research Council trial of streptomycin reported in 1948 [3]. The introduction to the trial pointed out that the natural history of pulmonary tuberculosis was so variable that "evidence of improvement or cure following the use of a new drug in a few cases cannot be accepted as proof of the effect of that drug." The introduction further pointed out that there had been only one report of an adequately controlled trial in tuberculosis and that was a trial of gold therapy [25]. This trial was negative and counteracted the exaggerated claims for gold treatment that had been made for over 15 years. Patients

entering the trial of streptomycin were restricted to those who were both unlikely to improve spontaneously and yet were likely to respond to an active chemotherapeutic agent. It was therefore decided that patients chosen had to have acute progressive bilateral tuberculosis; subjects were excluded if they had long-standing disease. The control treatment was to be bed rest and patients were excluded if they required pulmonary-collapse therapy.

The new feature of the trial was the randomisation of patients into control and treated groups. The report stated:

Determination of whether a patient would be treated by streptomycin and bed-rest (S case) or by bed-rest alone (C case) was made by reference to a statistical series based on random sampling numbers drawn up for each sex at each centre by Professor Bradford-Hill; the details of the series were unknown to any of the investigators or to the coordinator and were contained in a set of sealed envelopes, each bearing on the outside only the name of the hospital and a number. After acceptance of a patient by the panel, and before admission to the streptomycin trial, the appropriate numbered envelope was opened at the central office; the card inside indicated whether the patient was to be an S or a C case, and this information was then given to the medical officer of the centre.

Subsequent analysis showed that random allocation had led to the two groups being comparable at entry to the trial. After six months, 51 percent of the treated group showed considerable radiological improvement (radiographs were assessed without knowledge of the treatment group); only eight percent of the control group showed such improvement. Seven percent of the treated group were dead in six months as opposed to 27 percent of the control group.

The ethical considerations did not present a problem as bed rest was considered to be the only possible alternative treatment and only limited supplies of streptomycin were available. As not all cases could be given the new drug, it was reasonable and practicable to give it to a random half. The randomised controlled trial was therefore born just over 30 years ago and has since gone from strength to strength.

3. ETHICAL CONSIDERATIONS

3.1 DEFINITION

The *Concise Oxford Dictionary* defines ethics as the "science of morals." Glaser considered that a discussion of ethical problems should embrace both an assumption of right and wrong and a definition of how things are and not just how things should be [26]. Moreover, ethical problems concern the individual rather than the community. The community may benefit from the results of a trial but no individual should be asked to take an unreasonable risk to benefit the community. Problems arise when we are forced to consider what is reasonable.

Ethical considerations are not legal requirements, but the law may support an ethical stance. Lawyers usually consider precedents and determine the truth of matters by discussion. We can emulate this process for a definition of the term *reasonable*. At one extreme, it is obviously not ethical to force (or even request) a subject to take part in a dangerous study. Such trials were performed on non-Aryan prisoners in Nazi Germany. Subjects were exposed to extremes of temperature and trials of resuscitative techniques were employed. These experiments were obviously detrimental to and often fatal for the subject. Even if the experiments had revealed a resuscitative technique capable of saving the lives of many in the community, these trials were obviously unethical and the risks to the individual unreasonable. At the other extreme, every patient who agrees to take a medication must accept some risk. There are risks

to taking penicillin or aspirin but most subjects would be willing to face these risks in a controlled trial. This is an example of a reasonable or acceptable risk.

3.2 LEGAL CONSIDERATIONS

As Wade has pointed out [27], "Although the subject needs protection, the community needs knowledge." He considered how a subject should be indemnified if matters go wrong. The institution where the trial takes place must have a public liability insurance policy in case anything untoward happens to a subject as a result of negligence. With a new drug not in ordinary use, the policy may not cover such a contingency and, where applicable, the pharmaceutical company should agree to carry the risk. I also support Wade's idea that institutions should have no fault liability insurance so that subjects in trials may claim compensation for injury even when negligence does not occur. For example, a patient who experiences an adverse drug reaction while taking part in a trial could be recompensed.

3.3 DECLARATIONS ON MEDICAL ETHICS

3.3.1 The Nuremberg Code

Following the trials of Nazi war criminals, ten standards were laid down in 1947 [28].

1. The subject must give his or her voluntary consent, knowing the nature, direction, purpose, inconveniences, and hazards of the experiment.
2. The experiment should be necessary both in yielding fruitful results for the good of society and in the sense that the information cannot be gained without the experiment.
3. The anticipated results justify doing the experiment (see section 3.3.2. Clinical Research Combined with Professional Care and Nontherapeutic Clinical Research).
4. All unnecessary physical and mental suffering must be avoided (see The Use of Sham Operations, section 3.8.3).
5. There should be no a priori reason to believe that death or injury will occur.
6. The degree of risk shall not exceed the humanitarian importance of the problem (see section 3.1 and the discussion on reasonable).
7. Preparations should be made and adequate facilities provided against the remote possibility of adverse effects.
8. Those who conduct the experiment shall exercise the highest degree of skill and care and be scientifically qualified.
9. The subject must always be free to bring the experiment to an end.
10. The investigator must terminate the experiment if its continuation may be detrimental to the patient.

3.3.2 The declaration of Helsinki

The World Medical Association produced the following declaration [29], prefaced by binding the doctor with the words, "the health of my patient will be my first consideration."

I. *Basic Principles*

1. Clinical research must conform to the moral and scientific principles that justify research, and should be based on laboratory and animal experiments or other scientifically established facts.
2. Clinical research should be conducted only by scientifically qualified persons and under the supervision of a qualified medical man.
3. Clinical research cannot legitimately be carried out unless the importance of the objective is in proportion to the inherent risk to the subject.
4. Every clinical research project should be preceded by careful assessment of inherent risks in comparison to foreseeable benefits to the subject or to others.
5. Special caution should be exercised by the doctor in performing clinical research in which the personality of the patient is liable to be altered by drugs or experimental procedure.

II. *Clinical Research Combined with Professional Care*

1. In the treatment of the sick person the doctor must be free to use a new therapeutic measure if in his judgement it offers hope of saving life, re-establishing health, or alleviating suffering.
If at all possible, consistent with patient psychology, the doctor should obtain the patient's freely given consent after the patient has been given a full explanation. In case of legal incapacity consent should also be procured from the legal guardian; in case of physical incapacity the permission of the legal guardian replaces that of the patient.
2. The doctor can combine clinical research with professional care, the objective being the acquisition of new medical knowledge, only to the extent that clinical research is justified by its therapeutic value for the patient.

III. *Nontherapeutic Clinical Research*

1. In the purely scientific application of clinical research carried out on a human being it is the duty of the doctor to remain the protector of the life and health of that person on whom clinical research is being carried out.
2. The nature, the purpose, and the risk of clinical research must be explained to the subject by the doctor.
- 3a. Clinical research on a human being cannot be undertaken without his

free consent, after he has been fully informed; if he is legally incompetent the consent of the legal guardian should be procured.

- 3b. The subject of clinical research should be in such a mental, physical, and legal state as to be able to exercise fully his power of choice.
- 3c. Consent should as a rule be obtained in writing. However, the responsibility for clinical research always remains with the research worker; it never falls on the subject, even after consent is obtained.
- 4a. The investigator must respect the right of each individual to safeguard his personal integrity, especially if the subject is in a dependent relationship to the investigator.
- 4b. At any time during the course of clinical research the subject or his guardian should be free to withdraw permission for research to be continued. The investigator or the investigating team should discontinue the research if in his or their judgement it may, if continued, be harmful to the individual.

The Helsinki declaration clearly differentiated between the situation when the subject, usually a patient, can hope to benefit from the experiment and the situation where no such benefit can be expected.

Section II. 2 of the declaration stated that clinical research can be combined with professional care "only to the extent that clinical research is justified by its therapeutic value for the patient." This must be the overriding ethical consideration and the use of patients as volunteers for experiments not relevant to treatment presents great difficulties and will be discussed in section 3.5.

Sir Austin Bradford Hill has taken issue with two recommendations of the World Medical Association [30]. He found that there are experiments such as "in industrial psychology—which are not the prerogative, or even within the special competence, of the medically qualified," and he therefore objected to item I. 2, which insisted on the supervision of a qualified medical man. Hill also disagreed with the idea that the nature and purpose of the trial must be explained to the subject and stated "... I have no doubt whatever that there are circumstances in which the patient's consent to taking part in a controlled trial should be sought. I have equally no doubt that there are circumstances in which it need not—and even should not—be sought."

3.3.3 Bradford Hill's specific questions

Bradford Hill was unhappy with codes that deal in generalities and take no heed of "the enormously varying circumstances of clinical medicine" [30]. He stressed the necessity for "The close and careful consideration in the *specific circumstances of each proposed trial*" and formulated a series of questions to be answered for each trial.

1. Is the proposed treatment safe or, in other words, is it unlikely to do harm to the patient?
2. Can a new treatment ethically be withheld from any patients in the doctor's care?

Tuberculous meningitis was a universally rapidly fatal condition and when the first case reports revealed that streptomycin treatment had resulted in the patients' recovering, this fact was conclusive evidence of the effectiveness of the new treatment. It was then not ethical to perform a clinical trial of streptomycin in tuberculous meningitis. However, respiratory tuberculosis runs a more variable course and it was ethical to perform the randomised controlled trial of streptomycin in this condition. Moreover, only a limited amount of streptomycin was available at that time (1947) and as all cases could not be treated, it can be argued that it would be unethical not to have performed the trial.

3. What patients may be brought into a controlled trial and allocated randomly to different treatments?
4. Is it necessary to obtain the patient's consent to his inclusion in a controlled trial?
5. Is it ethical to use a placebo or dummy treatment?
6. Is it proper for the doctor not to know the treatment being administered to his patient?

3.3.4 Medical Research Council

A statement by the Medical Research Council (MRC) [31] gave two examples of when informed consent may not always be desirable. For example, when the patient has a possibly fatal illness without effective treatment being available, and second, when a placebo is employed. The MRC considered in 1964 whether any supervision of the conduct of controlled trials (or other experiments) was necessary and concluded "controlled clinical trials should always be planned and supervised by a group of investigators and never by an individual alone." The MRC report also suggested that no paper should be accepted for publication if there are any doubts about the ethical conduct of the study leading to the report.

3.4 RESEARCH ETHICAL COMMITTEES

In 1967 a committee appointed by the Royal College of Physicians of London suggested the formation of ethical committees consisting of "a group of doctors including those experienced in clinical investigation" [32]. By 1973 the functions and constitution of these committees had been formalised.

The final report made the following recommendations:

1. A Research Ethical Committee shall be a small committee set up solely to supervise the ethics of clinical research.

2. The medical members should be experienced clinicians with knowledge and experience of clinical research.
3. The Research Ethical Committee should have a lay member.
4. To remove any uncertainty about which procedures should be submitted to a Research Ethical Committee, all proposed research investigations in human beings should be submitted.
5. Whenever a research investigation was not expected or intended to benefit the individual patient a full explanation should be given and the patient should be free to decline to participate or to withdraw at any stage.
6. Whenever possible the consent of a patient should be obtained in the presence of a witness.
7. When there are circumstances in which it is genuinely inappropriate to inform a patient fully, it is the duty of the Research Ethical Committee to examine the situation with special care.
8. Particular care is needed if clinical investigation is proposed in children or mentally handicapped adults who cannot give informed consent. The parents or guardian should be consulted.
9. Particular care is needed if clinical investigation is proposed on a subject or patient who has any sort of dependent relationship to the investigator, for example, student, laboratory technician, or employee.

3.5 RANDOMISED CONTROLLED TRIALS WITHOUT POSSIBLE BENEFIT TO THE PARTICIPANT

Examples of these trials are provided by early drug studies in normal men and women and trials of drug interactions in patients on chronic treatment. Early drug trials in normal men are usually dose-finding experiments to assess the human counterpart of observations made in animals. They are not, initially, randomised controlled trials but slightly later studies may constitute a randomised trial of the new treatment (in the predetermined dose) versus an established drug (chapter 17).

In trials on patients, those on chronic treatment with one drug may be asked to take a second drug to assess the effect of the drugs in combination. This may be suggested when the second drug cannot be expected to benefit the patient. An example can be given of patients on long-term antihypertensive drugs who are also asked to take an antidepressant or antiinflammatory drug to assess whether or not the second drug worsens blood-pressure control.

For trials without possible therapeutic benefit for the individual, all subjects and patients must be true *volunteers*, receive *full information* about the study, give *written consent* (preferably in front of a witness), and *not receive an excessive reward*. If the subjects are paid a considerable amount they may be tempted to participate in a study, whereas without this remuneration they may refuse. This restriction does not exclude an allowance for fares, meals, and compensation for lost earnings as volunteers should not be expected to experience a financial loss.

3.6 WHO SHOULD PARTICIPATE IN TRIALS WITHOUT POSSIBLE THERAPEUTIC BENEFIT?

3.6.1 Employees of the pharmaceutical industry

Glaser outlined the case very clearly for using employees of the pharmaceutical industry [26].

Those who decide that a new substance can be safely tried in man should have enough confidence to take it themselves. If they will not take it themselves, they should not give it to others. Those who know the most about the substance and who are the most experienced scientists can make the best personal decisions about it and they are also the best able to observe their own subjective effects. Thus the first to take a new substance might be the research director, the medical director, the senior toxicologist, or advisers in pathology.

Glaser also considered that a volunteer's family doctor should be informed about the trial. This may ensure that trials that are unacceptable to general practitioners are not performed, and if there is some medical reason why an individual volunteer should not participate, then the investigator may be informed of this fact. Lastly, if the volunteer should become ill during or after the trial, then the general practitioner will be aware that the trial is in progress.

Volunteers must not be solicited from subordinates by their seniors. Only a comparatively junior person should perform this task and the supervisors should be told only who is suitable. Glaser reported "anyone unwilling is unsuitable" and children under the age of 14 and mental patients cannot volunteer.

3.6.2 Prisoners

Prisoners are used in medical experiments in the United States of America. The problem with this procedure is that a reduction in prison sentence may constitute an excessive reward and result in the subjects not being free volunteers. The report of a Committee appointed by the Governor of Illinois stated [33]:

A reduction in sentence in prison, if excessive or drastic, can amount to undue influence. If the sole motive of the prisoner is to contribute to human welfare, any reduction in sentence would be a reward. If the sole motive of the prisoner is to obtain a reduction in sentence, an excessive reduction of sentence which would exercise undue influence in obtaining the consent of prisoners to serve as subjects would be inconsistent with the principle of voluntary participation.

The committee considered the function of imprisonment, for example, whether this is to protect society or to reform the prisoner. The members discussed whether a prisoner would volunteer from good social consciousness or in a desire to reduce his sentence. In view of the latter incentive, the

committee concluded that a prisoner should not be allowed to volunteer if he is a habitual criminal or if he has committed a notorious or heinous crime. Presumably, the committee members were worried about having such a person released early.

The committee also concluded that any proposed reduction in sentence must not be excessive. Glaser [26] also worried that the incentives for prisoners may be too high. He considered the possibility of prisoners getting privileges for participation and even that the relief of boredom might prove a great incentive, possibly a coercion inconsistent with voluntary participation.

3.6.3 Patients

Patients are the ultimate beneficiaries of advances in medical care and Claude Bernard considered it their duty to assist with research. However, should the individual patient in the trial be the possible beneficiary or should the benefit go to other patients with different conditions? If we wish to assess the interaction between an antihypertensive and an antiinflammatory drug, we may ask any hypertensive patient or patients with both hypertension and arthritis to cooperate. In the latter instance, the treatment is relevant to the patient's condition and the patient may benefit. However, when the patient has hypertension alone, he must be considered as a normal volunteer and great care must be taken that the doctor-patient relationship is not used to exert too much pressure on the patient to participate. The patient must not volunteer from a sense of gratitude or in the hope of better medical attention and it is a wise precaution for the doctor treating the patients to ask them to discuss taking part in a trial with another colleague. The doctor undertaking the usual treatment should make it clear that the patient's failure to participate in a trial will not affect his usual medical care in any way.

3.7 INFORMED CONSENT

3.7.1 Information for the patient or subject

The patient or subject should be fully informed of the nature of the trial: that is, the number of investigations and visits required and the duration of the trial. The objectives of the trial should be stated, provided such statements are compatible with the usual doctor-patient relationship. It may be unethical to give full information to patients with, say, cancer, either when the diagnosis cannot be revealed or when it is not in the patients' best interest to describe the inadequacies of available treatment. However, in most instances the patients can be given all the relevant information and should be told when a placebo (dummy) treatment is to be employed in the trial. In conclusion, the patients should be informed of the following, in writing, and preferably in the presence of a witness:

1. the nature of the treatments being compared
2. the objectives of the trial

3. the duration of the trial
4. what the trial involves for the patients (number of visits, investigations, et cetera)
5. possible benefits to be derived from the treatments
6. possible hazards of the treatments
7. what to do if the patients become unwell, run out of tablets, et cetera

3.7.2 Written consent

Written consent should be obtained; otherwise, there can be no proof that consent was given and such evidence may be necessary in a court of law. The patients should be asked to sign a document giving the full information discussed in section 3.7.1 and including a declaration similar to the following: "I ———, have read the above description of the trial and agree to take part. I understand that I may withdraw my co-operation at any stage should I so wish." The patients therefore sign to say they have been informed about the trial and have agreed to take part. Some authorities may insist that the declaration be signed in the presence of a witness. This is desirable in all volunteer studies but many researchers would not insist on the presence of a witness when the trial is of possible therapeutic benefit to the individual involved.

3.7.3 How to avoid asking consent of some of the patients

When a new experimental treatment is to be compared with an acceptable routine treatment it can be argued that only the patients receiving the new treatment need give consent. The usual trial design requires consent to be obtained prior to randomisation but Zelen has suggested that randomisation can precede informed consent so that only those allocated to the new treatment are asked to consent [34]. There may well be a circumstance in which this strategy is desirable; however, the approach is impossible in double-blind or single-blind trials. Moreover, Zelen's suggestion may be unsatisfactory if the new treatment proves to represent an important new advance: the patients who have benefitted will have consented to take part but not those who have fared badly. Most important, however, is the fact that those who do not wish to take part will fail to receive the trial treatment in the consent group but will receive it in the no consent group. Patients who do not receive the treatment cannot be excluded; otherwise, the two groups may be dissimilar for important characteristics and one major purpose of randomisation will be lost. Analysis has to be conducted on all randomised patients on the intention-to-treat principle (section 15.7). However, the effect of the new treatment may be diluted by the results in patients who do not receive this treatment.

3.8 PLACEBO TREATMENT

The reasons for using placebo drugs and the methods for using such drugs are discussed in section 8.10. There are two circumstances where it is ethical to employ placebo medication; when no effective treatment is available for a

particular condition and when, if such treatment is available, it can safely be withheld for a certain period.

3.8.1 No effective treatment has been identified

A placebo cannot be employed if there is definite evidence that withholding standard treatment would be detrimental to the patient's health. Beecher [35] discussed a trial of treating streptococcal respiratory infections in which placebo was given to 109 men while benzathine penicillin G was given to the others. No patient treated with penicillin developed either rheumatic fever or acute nephritis. However, three patients developed these complications when given placebo. Beecher considered that at the time the trial was performed it was known that penicillin prevented rheumatic fever, and therefore the use of a placebo was unethical.

3.8.2 Use of a placebo for a short period when active treatment is known to be required

We can agree that it is unethical to withhold necessary treatment. However, a placebo may have a powerful pain-relieving effect and constitute acceptable treatment under certain circumstances. Beecher [36] reviewed the effects of placebo in severe postoperative wound pain. Four studies used an injection of saline as a placebo and satisfactory pain relief was achieved in about a third of patients. Similarly, placebo treatment produced relief from angina pectoris in a similar proportion of patients (section 8.10).

In patients with severe postoperative pain, we can argue that placebos should not be used, as active drugs such as morphine are available. But arguments for using a placebo in this situation can be advanced. A proportion of patients achieve pain relief from placebo and they do so without the adverse effects associated with active pharmacological agents. Although it would be unethical to withhold an active analgesic for a prolonged period, administration may be delayed for a short interval, say, 15–30 minutes following a placebo injection. If pain relief is not achieved at the end of this period, active treatment can then be given. Many patients will agree to wait a short period to help in evaluating a new drug. Care has to be taken in a double-blind study of a new drug against placebo that there is no likelihood of an adverse effect if the new drug is ineffective and is closely followed by an active drug such as morphine.

With less severe degrees of pain the ethical problems are reduced and it is even more necessary to employ a placebo to assess a possibly active mild analgesic. Many patients will respond adequately to a placebo and not all patients receiving active treatment will experience pain relief. For example, if the improvement rate is 33 percent with placebo and 50 percent with an active treatment, the use of a placebo is necessary to confirm a superior effect for the active compound. Without knowledge of the placebo response rate, a 50 percent result could be a nonspecific response to an inactive compound.

Placebo treatment has also been employed for short periods in trials in

chronic diseases when active treatment is known to be required in the long term. An example is provided in hypertension where placebo treatment in the long term is justified for patients with mild hypertension as the benefits of active treatment have not been established. However, active treatment is known to be beneficial in preventing cerebrovascular events in young or middle-aged patients with moderate or severe hypertension. Yet placebos are prescribed when immediate treatment for heart failure, renal failure, or malignant hypertension is not required. Placebo treatment is traditionally employed in two broad circumstances.

1. *The first is when the patient has not received antihypertensive treatment in the past.* Antihypertensive treatment is not usually started the first time the patient sees a doctor. The physician may wish to confirm that the blood pressure is elevated on a second or third occasion and may require certain investigations to be completed before commencing treatment. It is therefore reasonable to give placebo treatment during this period of observation and possibly to extend the interval to, say, four to six weeks. It must be appreciated, however, that the preventive effects of antihypertensive treatment are being denied the patient during this period. Although the risk of a cerebrovascular event occurring during a short interval is low, it still constitutes an ethical problem. The theoretical risk can be calculated from the Veterans Administration trial of antihypertensive treatment [37]. Male patients less than 60 years old with an initial diastolic blood pressure of 105–114 mm Hg experienced, on average, a 0.058 chance of a mortal or morbid cardiovascular event per year of placebo treatment. If they were given active treatment the chance of a cardiovascular event was 0.023 per year. The excess risk of being on a placebo was therefore 0.035 events per year and the probability of a mishap with placebo treatment may be 0.003 per month. If 220 such patients receive six weeks of placebo treatment in a trial, the investigator may expect one adverse event. These events include stroke, myocardial infarction, heart failure, and the retinal changes of accelerated hypertension. Many trials have subjected patients to a one in 220 risk of one of these events.

2. *The second circumstance is when placebo treatment interrupts a period of active treatment.* The risks of taking placebo treatment may be the same when a period of placebo treatment interrupts active treatment as when it precedes it, and many patients on treatment have been entered into trials that incorporate a period of placebo administration. Again, for a six-week period we must ask if the patient will accept a one in 220 chance of an adverse cardiovascular event.

3.8.3 The use of sham operations

The immediate reaction among physicians is to consider all sham operations in man as unethical. Bradford Hill considered that it would not have been reasonable to use placebo injections as a control in the Medical Research Council

(MRC) trial of streptomycin [3]. In this trial the control patients would have suffered a considerable amount of discomfort from repeated injections of placebo. Admittedly, if injections alone can have a life-prolonging effect independent of the substance injected, then the MRC trial was biased in favour of streptomycin. However, such a strong placebo effect was unlikely and could not justify inflicting so much discomfort on the control patients.

Before dismissing sham operations, however, we must consider a trial discussed by Beecher [38]:

In 1939 it was suggested, in Italy, that the pain of angina pectoris could be greatly lessened by ligation of the internal mammary arteries. Eventually this suggestion was adopted in the United States and quite spectacularly favourable results were obtained. Not only were the objective results impressive, the patients said they felt better and the objective evidence supported this: there was great reduction in the number of nitroglycerin tablets taken, and exercise tolerance was greatly increased. Several individuals [39–42] began to wonder if this might not be a placebo effect. They therefore went to their patients, explained the situation, and told them they would like to carry out a study in which the patients would not know what had been done, nor would the observers know until the study was completed. They told their patients that half of them would have the internal mammary arteries exposed and ligated and the other half would simply have them exposed, but not ligated. These studies were carried out . . . ligation had no real effect beyond that of a placebo effect.

Beecher thus argued very persuasively that a sham operation can be ethical even though the control patients suffered an anaesthetic and much discomfort. Many patients would not agree to take part in such a trial.

3.9 SELECTION (EXCLUSION OR INCLUSION) CRITERIA

Exclusion and inclusion criteria are the two sides of the same coin. A trial confined to young patients may be said to *exclude* patients above the age of 60 or *include* only patients below the age of 60.

Two objectives are met by using these criteria. First, only those patients who are intended for study are entered into the trial. The results of the trial are then only valid for a similar group of patients; this concept will be discussed further in chapter 5. The second reason for having selection criteria is ethical. Patients must be excluded from a trial if inclusion in the trial may produce adverse consequences for them.

Table 3–1 gives the selection criteria for a placebo-controlled trial of antihypertensive treatment in the elderly, being conducted by the European Working Party on Hypertension in the Elderly (EWPHE) [43]. The selection criteria are rearranged into criteria defining the group of patients to be studied, and criteria excluding patients from the study who should not be included for ethical reasons. The trial involves the random allocation of patients either to five years of active treatment or five years of placebo treatment. The selection criteria therefore exclude patients who should not receive a placebo for five years. At

Table 3-1. Selection Criteria for the European Working Party on Hypertension in the Elderly (EWPHE) Trial of Antihypertensive Treatment.

Selection criteria defining the group of patients to be studied

1. Aged more than 60.
2. Systolic blood pressure (on placebo) above 160 mm Hg.
3. Diastolic blood pressure (on placebo) above 90 mm Hg.
4. Patients give their informed consent.
5. Regular follow-up possible.
6. Compliant with medication as assessed by pill count.
7. No reason to suspect secondary hypertension.
8. No severe life-threatening diseases unrelated to hypertension (e.g., carcinoma).

Selection criteria included for ethical reasons

9. Systolic blood pressure (on placebo) not above 239 mm Hg.
10. Diastolic blood pressure (on placebo) not above 119 mm Hg.
11. No history of accelerated or malignant hypertension.
12. No congestive heart failure.
13. No severe renal failure (serum creatinine > 2.5 mg %).
14. No previous history of a haemorrhagic stroke or hypertensive encephalopathy.
15. No history of dissecting aneurysm.
16. No previous history of gout or serum uric acid > 10 mg %.
17. No acute hepatitis or active cirrhosis.

the time of initiating the trial (and at the time of writing), there was little or no evidence that the elderly hypertensive patient would benefit from antihypertensive treatment. However, it was considered undesirable to include patients with very high levels of blood pressure (criteria 9-10). Similarly, hypertensive patients known to require treatment were excluded, for example, those with accelerated or malignant hypertension (criterion 11), those with congestive heart failure, and those with conditions that would possibly benefit from treatment, such as patients with renal impairment (criterion 13) or those who had previously suffered a haemorrhagic stroke (criterion 14).

Selection criteria must exclude not only those patients who would suffer from placebo treatment, but also those known to be adversely affected by the active treatment. Criteria 16-17 exclude patients who may be adversely affected by the active treatments employed in the trial (hydrochlorothiazide with triamterene; and methyldopa).

3.10 WITHDRAWALS FROM THE TRIAL

If a patient with a certain condition cannot enter the trial for ethical reasons, then he should be withdrawn from the trial if he develops the condition. Withdrawal criteria should therefore be the same as exclusion criteria. Table 3-2 gives the withdrawal criteria for the EWPHE trial. Criteria 1-3 are end points for the trial and not ethical considerations. Criteria 4-10 correlate with the exclusion criteria given in table 3-1. The criteria in the two tables are cross-referenced in table 3-2. Selection criteria 16-17 (table 3-1) do not have their counterparts in the withdrawal criteria in table 3-2, as the development of gout may lead to the discontinuation of diuretic treatment, but the patient

Table 3-2. Withdrawal Criteria for the European Working Party on Hypertension in the Elderly (EWPHE) trial.

Withdrawal criteria which are end-points for the trial and not ethical considerations.

1. Completion of five years follow-up.
2. No follow-up for more than six months.
3. No trial treatment for more than three months.

Withdrawal criteria for ethical reasons.

(In parentheses—the corresponding numbers for the selection criteria)

4. Systolic blood pressure rising by 40 mm Hg or exceeding 250 mm Hg on three visits (9).
5. Diastolic blood pressure rising by 20 mm Hg or exceeding 130 mm Hg on three visits (10).
6. Development of accelerated or malignant hypertension (11).
7. Development of congestive heart failure (12).
8. Serum creatinine increasing by 100% or above 3.9 mg % on two occasions (13).
9. Development of cerebral or subarachnoid haemorrhage or hypertensive encephalopathy (14).
10. Development of a dissecting aneurysm (15).
11. Voltage criteria for a 30% increase in left ventricular hypertrophy as assessed from the electrocardiogram.
12. A 20% increase in the cardio-thoracic ratio as measured on a chest radiograph.
13. Any reason why continuation in the trial would be detrimental to the patient's interest.

may continue in the trial taking methyldopa. Similarly, the development of liver disease may lead to stopping methyldopa, with the patient continuing to take a diuretic and remaining in the trial. Withdrawal criteria 11 and 12 have no counterpart in the selection criteria but indicate that the patient is not progressing satisfactorily.

However carefully a trial is designed, and even after the completion of a pilot trial, there will still be patients whose continuation in the trial would be against their future well-being. Criterion 13 allows for these unforeseen contingencies and is a necessary statement in any trial protocol.

3.11 DECISION RULES FOR STOPPING THE WHOLE TRIAL

In a short-term trial, the patients are usually entered into the trial quickly and the trial completed before the results are analysed. The exception to this rule is the sequential trial where a decision is made whether or not to continue with the trial as the individual results become available (section 11.7). When patients are followed for several years or when recruitment persists for many years, the opportunity exists for interim analyses to be made. For ethical reasons the trial must be terminated if an interim analysis demonstrates a statistically significant and important adverse effect of treatment or a significant benefit from treatment. If interim analyses fail to reach these end points the trial will be terminated when the intended number of patients has entered the trial or the trial participants and organisers run out of time or money (section 14.6).

3.11.1 Problems with significance testing in interim analyses

Care has to be taken that the overall level of significance of a trial is not reduced by the repeated analysis of results. These interim analyses are some-

times termed *repeated looks*. If a statistical test is repeated on several occasions on increasing data and five percent is taken as the level of significance to be achieved, then after the first test the probability of a falsely positive result is five percent. After two tests this probability rises to almost ten percent. After 13 tests, the chances of a falsely positive result is almost 50 percent.

McPherson [44] has calculated that ten interim analyses with a decision rule to stop the trial if the level of significance exceeds one percent is equivalent to an overall level of significance of five percent (see section 10.7.3). In other words, if the trial must be stopped for an adverse effect significant at the five percent level and ten interim analyses are planned, then the result of an interim analysis must be significant at the one percent level to stop the trial.

3.11.2 Terminating the trial when an adverse effect of treatment is observed

Treatment with either conjugated oestrogens or dextrothyroxine in the Coronary Drug Project Trial [45] had to be terminated owing to the adverse effects of these drugs. Similarly the University Group Diabetes Project (UGDP) trial was stopped owing to the adverse effects of phenformin and tolbutamide [46] (see section 19.6). In both trials the groups treated with certain drugs fared significantly worse than the placebo-treated groups and the trials of these active treatments were terminated.

3.11.3 Terminating the trial when a statistically significant benefit is observed

The Veterans Administration trial of antihypertensive medication provides a good example of a trial's being terminated when an interim analysis provides evidence of a benefit from treatment. Patients were entered into the trial when the diastolic blood pressure ranged from 90–124 mm Hg while taking a placebo and they were randomly allocated to receive either active or placebo treatment. After an average of 18 months' follow-up the trial was stopped for patients with an initial diastolic blood pressure greater than 114 mm Hg [47], as the patients receiving active treatment had fewer cardiovascular events than those receiving placebo ($P < 0.001$). The trial was continued for patients with an initial diastolic blood pressure of 90–114 mm Hg [37]. Another interesting example comes from the Anturane Reinfarction Trial [48] (section 19.1). In this trial a significant benefit from treatment was observed in an interim analysis and recruitment to the trial was stopped. However, the patients already in this double-blind trial were advised of the results of the interim analysis and asked to continue in the trial. Nearly all agreed to continue and the final analysis showed a similar benefit.

3.12 TRIALS TO DETECT TOXICITY

It is unethical to design a trial to detect toxicity. However, as discussed in section 18.2, large long-term trials have resulted in the detection of unexpected treatment toxicity, although some large trials have failed to detect rare adverse

drug effects. Large trials are designed to estimate the efficacy of treatment but careful attention must be paid to possible toxicity (chapter 18).

3.12.1 When may a trial to detect both efficacy and toxicity be desirable?

When a single trial reports a benefit from treatment, it is often desirable to repeat the trial and ensure that the benefits can be demonstrated for different patients and on another occasion. However, when a trial detects toxicity, it is ethically impossible, although scientifically desirable, to conduct a trial to confirm an adverse effect of treatment. If there are doubts about whether or not there was any serious toxicity in the earlier trial, and the efficacy of the treatment is thought to be high, then possibly a second trial of benefit can be mounted. The question is of some practical importance. For example, in the University Group Diabetes Program Trial [46] oral hypoglycaemic agents were associated with an increase in cardiovascular mortality. However, when a patient cannot adhere to a diet these drugs may relieve the symptoms of hyperglycaemia. It may be reasonable to reassess efficacy in these patients and in view of the criticism levelled at this particular trial (section 19.6), the trial could be repeated. I would be reluctant to take part in a trial where toxicity may be a disadvantage *not* counteracted by important gains from therapy.

3.13 CONCLUSIONS

This chapter reviews the ethical requirements in the design and conduct of clinical trials. Declarations of ethical principles have been reviewed and the place of research ethical committees considered. Emphasis was placed on the importance of obtaining informed written consent and a distinction has been drawn between trials of possible benefit to the participant and trials involving volunteers who cannot expect an improvement in their health from participating in the trial. The investigator must remain convinced that none of the available treatments offer a clear advantage and this is especially important when placebo treatment is to be employed. Provided the investigator is genuinely in doubt as to the best treatment, he can explain the situation to potential participants and ask them to enter the trial. He may even ask himself the standard question, "Would I allow a member of my family to enter the trial?" Even if he can answer yes to this question, the public must be protected from a small proportion of eccentric enthusiasts; research ethical committees should provide this safeguard.

Large trials should incorporate an ethical committee in the administration that is independent of the investigators and rules on whether any observed toxicity is acceptable, when the trial should be terminated, and whether to make any changes in the criteria for entry or withdrawal from the trial.

If a trial shows one treatment to be superior, patients who received the inferior treatment may have suffered as a consequence. Trial designs that limit this problem are discussed in sections 11.7 and 11.8 and the ethical disadvantages of randomised controlled trials are summarised in chapter 20.

4. THE OBJECTIVES OF A RANDOMISED CONTROLLED TRIAL

A trial may be conceived to test more than one hypothesis but it is good practice and usually essential when calculating the numbers required for a trial to determine one major objective. For example, an investigator may be interested in a trial of a new antihypertensive drug in elderly patients. The major objective could be either to demonstrate the efficacy of the drug in lowering blood pressure or in preventing cardiovascular deaths. The first objective could be answered in a few patients studied for six months, but the second objective would require the study of hundreds of patients over many years (chapter 10). In order to calculate the numbers required for a trial, the major objective has to be identified and the smallest effect of treatment to be detected must be defined.

4.1 IDENTIFICATION OF THE MAJOR OBJECTIVE

The major objective will involve the detection of a change in a particular end point. In the first example given earlier, where the effect on blood pressure has to be determined, the end point of interest could be diastolic, systolic, or mean pressure. If the effect on mortality or morbidity has to be determined, total mortality, total cardiovascular mortality, stroke mortality, total cardiovascular events (either fatal or nonfatal), or stroke events could constitute the major end point of primary interest. The investigators must determine this end point at the outset of the trial and also decide the amount by which it should change. In the definition of the major objective the investigators must take into account

its importance, the likelihood that it will be achieved by treatment, and the case of measurement of the chosen end point.

4.1.1 The importance of the objective

Systolic blood pressure may be easier to measure than diastolic pressure but the investigators may decide that diastolic pressure is more important in determining the future health of the patient. Similarly, a death from myocardial infarction may be more important than the occurrence of an infarct from which the patient recovers. Mortality is usually a more important end point than morbidity and total mortality is a clearer measure of outcome than mortality from one specific cause. It must be noted that a treatment may reduce one cause of death and increase another.

4.1.2 The likelihood that the objective will be achieved by treatment

An antihypertensive drug may reduce both stroke mortality and total mortality. However, the proportional reduction in deaths may be expected to be greater with stroke mortality. A trial with stroke mortality as its end point is therefore likely to reach a conclusion more quickly than a trial to detect a reduction in total mortality. The end point of stroke mortality is to be preferred in this example for ethical reasons and efficiency. However, total mortality has also to be considered and this is discussed under the decision rules for stopping a trial (section 14.6).

4.1.3 Ease of measurement of the end point

It may be easier to measure systolic blood pressure than diastolic pressure and this measurement may have greater repeatability. Similarly, the fact of death is easier to determine than a particular cause of death and a cardiovascular event may be particularly difficult to ascertain. If the patient dies it has to be decided whether or not sudden death should be regarded as a cardiovascular death, and, if so, how quickly must the patient die to be considered as a sudden death. Also, if the patient survives a myocardial infarction, the diagnostic electrocardiographic or enzyme changes have to be agreed in advance.

4.2 WHAT CHANGE IN THE END POINT MUST BE DETECTED?

We must distinguish between biological and statistical significance, consider whether the end point is a continuously distributed or qualitative variable, and, if qualitative, whether or not the end point occurs frequently or rarely.

4.2.1 The distinction between biological and statistical significance

The distinction has to be made between what is statistically significant and what is biologically important. It may be observed that an anti-anxiety drug lowers systolic blood pressure by, say, 2 mm Hg and this result, given a large number of patients, could be highly statistically significant. However, the biological importance of the result would be small and the drug would not be

Table 4-1. Average results and standard deviations for the results of a survey on 634 London Civil Servants aged 35-64, 34% of whom were female.

	Mean	Standard Deviation	Biologically Important Change	
Systolic blood pressure (mm Hg)	133	20	- 10	(- 8%)
Diastolic blood pressure (mm Hg)	82	13	- 8	(- 10%)
Blood haemoglobin (gm/100 ml)	14.5	1.1	+ 2	(+ 14%)
Blood glucose (mmol/l)	5.5	1.4	- 1	(- 18%)
Serum cholesterol (mmol/l)	6.3	1.0	- 0.6	(- 10%)
Serum urate (mmol/l)	0.32	0.07	- 0.05	(- 16%)
Serum creatinine (mmol/l)	93	15	- 15	(- 16%)

used as an antihypertensive agent. In hypertension, a drug is useful if it lowers systolic blood pressure by more than 10 mm Hg. The objective of the trial would therefore be to test the hypothesis that the drug lowers systolic blood pressure by 10 mm Hg (or conversely, the null hypothesis that the drug does not do so).

4.2.2 Changes in continuously distributed variables

What sort of changes in continuously distributed variables are of biological importance? Table 4-1 gives some biochemical and other results from a screening of London civil servants [49]. The table also gives some very arbitrary suggestions for changes that could possibly be produced by treatment and be considered biologically important. The suggested changes are of the order of ten and 20 percent or about one standard deviation. For example, a reduction in systolic blood pressure of 10 mm Hg; an increase in haemoglobin of 2 gm/100 ml and a reduction in blood glucose of 1 mmol/liter could all be considered biologically important.

4.2.3 Changes in qualitative end points

When defining a change in proportion, we may have more difficulty in identifying a biologically important effect. For example, if we are considering a reduction in mortality it could be argued that any reduction, however small, is important. On the other hand the cost and adverse side effects of treatment may negate small benefits. In addition, with treatment to prevent an uncommon event, small benefits become less acceptable. It has been suggested that ten individuals with mild hypertension would have to take antihypertensive medication for 20 years to reach an even chance of avoiding one cardiovascular event. If the probability of an event is low, a treatment used for prevention must be highly effective, whereas a patient with an incurable illness will be interested in a trial treatment that offers a cure only in a small percentage of

patients. We shall consider the infrequent and frequent end point in more detail.

4.2.3.1 The end point occurs infrequently

Cardiovascular events, though not rare in the general population and common in patients with a previous history of cardiovascular disease, may occur very infrequently during a controlled trial. A trial of secondary prevention is intended to prevent a recurrence of a condition and is to be contrasted with a primary prevention trial intended to prevent the condition initially. Even with a secondary prevention trial of myocardial infarction, fewer than 15 percent of patients who leave hospital will die over a one-year period. For the purpose of this discussion such events will be considered infrequent. Table 4-2 illustrates some trials that have suggested a benefit in preventing cardiovascular disease; indicates whether or not the authors of the trials have suggested the treatment be adopted; and considers whether the benefits have been accepted and the findings implemented by the medical community at large. The trials included in the table include the Veterans Administration Cooperative Study Group on Antihypertensive agents [37, 47] and trials designed to test sulphinpyrazone (Anturane), aspirin, clofibrate, and anticoagulants in the primary and secondary prevention of ischaemic heart disease. The 93 percent and 77 percent reductions in morbid or mortal events in the Veterans Administration antihypertensive agents trial have been accepted and acted upon by the medical profession. However, when treatment in this same trial produced only a 33 percent reduction in cardiovascular events, the results were not widely accepted and many physicians do not treat a diastolic pressure of 90-104 mm Hg. Also, although it can be argued strongly that a 20 percent reduction in deaths is worthwhile, the medical profession has not considered such benefits warrant the expense and difficulty of long-term treatment. Anticoagulant treatment has therefore fallen from favour and the use of sulphinpyrazone (Anturane), aspirin, and clofibrate treatment has not been widely accepted. It must be admitted that inconsistencies in the data have not helped. For example, anticoagulants have little effect on the death rate in women and treatment can only be provided for men. Similarly, total deaths were increased by clofibrate and it would not be acceptable to employ this treatment.

Clinicians tend to ignore small benefits, and when the event rate for a disease is low the patient may also be unwilling to take treatment to reduce a small risk by only, say, 20 percent. The patient will be more interested in therapy that almost guarantees freedom from the disease.

4.2.3.2 When the end point for the trial is very frequent

If a condition gives rise to a high mortality, as with certain cancers, both the doctor and patient may be interested in a drug that reduces mortality by less than 20 percent. A patient faced with no hope of recovery may be pleased to accept a one in five chance or less of survival.

Table 4-2. The results of several large trials to detect a reduction in cardiovascular disease. The benefits are listed, together with the authors' recommendations and subsequent use of the treatments.

Trial	End point	Reduction	Did authors recommend treatment?	Has the treatment been subsequently widely used?
<i>Primary prevention</i>				
Veterans Administration Co-operative study group on Antihypertensive agents [37, 47]	Morbid or mortal events			
Diastolic 115-129 mm Hg		93%	Yes	Yes
Diastolic 105-114 mm Hg		77%	Yes	Yes
Diastolic 90-104 mm Hg		33%	Yes	?
Clofibrate [50]	IHD incidence	20%	No	No
	Nonfatal MI	25%	No	No
	Fatal IHD	(8% increase)	No	—
	All deaths	(22% increase)	No	—
<i>Secondary prevention of myocardial infarction</i>				
Anturane [48]	Sudden deaths (males)	43%	Yes	?
Aspirin [51]	Total mortality	25%	No	No
Anticoagulants [52]	All deaths (males)	20%	Not clear	No
	All deaths (females)	8%	No	No

IHD = Ischaemic heart disease; MI = Myocardial infarction.

4.2.4 Other considerations necessary for the definition of a major objective

The major objective must not only include a statement on the size of the treatment effect that must be detected but also the statistical significance with which it should be reported (section 10.2), and if the treatment effect is near zero, the confidence with which the specified treatment effect can be excluded (section 10.3).

4.3 SPECIFICATION OF MINOR OBJECTIVES

Minor objectives are subsidiary to the main purpose of the trial and are not used to calculate the number required for the trial. Minor objectives may be specified in order to fully interpret the conclusions originating from the trial. For example, it would make little sense to mount a long-term trial to assess the reduction in stroke incidence (main objective) without assessing the adverse effects of treatment. Minor objectives may be required to assess the trial results or may be only loosely connected with the main purpose of the trial. It is desirable to answer additional questions where possible, provided the smooth running of the trial is not jeopardised. In the previous example a drug treatment may be employed to reduce the incidence of stroke and the opportunity may be taken to examine the effect of the drug on blood fats. We shall discuss a few more examples of minor objectives and the factors limiting their investigation.

4.3.1 Examples of the minor objectives that can be defined in a long-term trial

Table 4-3 lists some minor objectives that may be formulated during a long-term trial of drug treatment to reduce mortality. The major objective is given first and includes a statement of the size of the effect to be detected, the level of

Table 4-3. Possible objectives that may be defined during a trial of drug treatment versus placebo in reducing mortality from a chronic disease.

<i>Major objective</i>
1. Does the drug treatment reduce mortality from disease V by X% (with a significance of Y and a power of Z)?
<i>Essential minor objectives</i>
1. Does the drug treatment reduce or increase total mortality?
2. To determine any serious adverse effects of drug treatment.
<i>Other minor objectives</i>
1. To determine the natural history of the disease while on placebo.
2. To identify any biochemical effects of drug treatment.
3. To determine the interrelationships between the condition under treatment and other diseases.
4. To identify any symptomatic side effects of drug treatment.
5. Objectives unrelated or indirectly related to the condition being treated (e.g., factors influencing compliance with therapy, . . .).

significance to be achieved and, in the event of the effect not being demonstrated, the confidence with which it is excluded (power).

The essential minor objectives in this type of trial will include an assessment of both total mortality and any serious but nonfatal adverse effects of treatment. The objectives can be illustrated by considering again the European Working Party trial of Hypertension in the Elderly (EWPHE) [43]. The major objective was defined as a reduction in stroke events (mortal plus morbid events) of 50 percent, with a level of significance of 5 percent and a power of 90 percent. Certain minor objectives listed in table 4-3 have been studied and resulted in a description of the natural history of untreated hypertension in the elderly and an examination of the changes in cardiac and renal function with increasing age in these subjects [53]. In addition, the biochemical changes with active treatment were estimated and a reduced glucose tolerance in the group treated with a diuretic reported [54]. Other minor objectives listed in table 4-3 include the interrelationship between the condition under treatment and other diseases, the side effects of treatment, and items of general interest such as the factors influencing compliance with treatment or default from follow-up.

4.3.2 Factors limiting the investigation of minor objectives

The greater the number of objectives, the greater will be the complexity of conducting the trial. A trial with many minor objectives may require repeated biochemical and other investigations and impose a great burden on the investigators and the subjects, both of whom may be required to make more time available for the trial. Additional expenses will also be incurred and clerical duties increased. However, without answering certain important questions, it may be difficult to assess the results of a trial.

4.4 CONCLUSIONS

It is very important to specify a major objective for a trial. This must include the most important end point to be measured, the size of any treatment effect that has to be determined, the significance with which the effect must be observed, and the power of the trial in the event of a nonsignificant result being reported. Armed with this information and an estimate of the variance of the end point under consideration, the numbers required for the trial can be calculated (chapter 10).

Minor objectives may have to be specified in order to fully appreciate the outcome of the trial. Some will be necessary for this purpose and others incidental to the main objective of the trial. How many minor objectives are defined will depend on the resources available, but the collection of a large amount of information during the course of a trial may hinder its successful completion.

5. VALIDITY OF THE RESULTS

Having defined the objectives for a trial, we must consider the validity of any results we may obtain. One dictionary definition of validity is "so executed etc. as to have binding force" [55]. A trial result may not have binding force when executed incorrectly or when the trial provides a certain result but the interpretation of that result is incorrect. An incorrect interpretation can arise when the trial reveals a particular result and the investigator jumps to a further conclusion or when he decides that the same result will be true in different subjects.

5.1 THE TRIAL SHOWS THAT THEREFORE THIS MUST BE TRUE

We shall consider a hypothetical trial that provides an example of a non sequitur. The trial is designed to determine whether or not a particular pharmaceutical agent can stop cigarette smokers from indulging in their habit. The trial shows that the drug stops smoking in a significant proportion of smokers. However, the authors believe that smoking causes heart disease and conclude that the drug will reduce total cardiovascular mortality. This does not follow because when a causative factor is removed the previous adverse effects may persist. A different trial is required to show that stopping smoking prevents cardiovascular death and, in fact, one trial of antismoking advice showed that stopping smoking was associated with a reduction in respiratory symptoms but not total mortality [56].

Similar examples in the cardiovascular field are provided by trials showing

that antihypertensive drugs lower blood pressure. Such trials do *not* prove that the drugs reduce stroke mortality. The results of a trial must not be extrapolated beyond the observations. It is not valid to conclude that a treatment has wider effects than those observed.

5.2 THE VALIDITY OF THE RESULTS FOR SUBJECTS OF A DIFFERENT AGE, SEX, OR RACE

The response to a treatment may vary between men and women, alter with age, and sometimes differ among the races. We cannot assume that the results of a trial including, for example, only young men are applicable to elderly women as well. This mistake commonly arises when trials are conducted on laboratory staff (these persons tending to be young and male); an example of this problem is often provided when the clinical pharmacology of a new drug is determined on volunteers.

5.3 THE VALIDITY OF THE RESULTS FOR OTHER PERSONS OF THE SAME AGE, SEX, AND RACE

Even after allowing for age, sex, and race the selection of subjects into a trial makes it difficult to apply the results to even a superficially similar group. With patients, selection may favour those with either a particularly severe or mild form of the disease. Also, patients taking part in clinical trials are often more compliant with therapeutic advice than patients subsequently offered the treatment. Lastly, a trial result may occasionally depend heavily on a subgroup of subjects.

5.3.1 The severity of the disease

The Veterans Administration Co-operative Study Group on Antihypertensive Agents trial [47] admitted only men who were patients attending Veterans Administration hospitals; these men appear to have had moderate to severe hypertension. First, the level of diastolic blood pressure leading to entry to the trial was from 90–129 mm Hg after *four to six days resting in hospital*. This level of hypertension had also to be observed in an outpatient department after a two-to-four-month period of placebo medication, at which time the pressure was designated the untreated pressure. In addition the diastolic blood pressures were measured at the point of disappearance of the Korotkoff sounds, not the point of muffling of these sounds.

A patient in a doctor's office with a similar level of untreated pressure may not be comparable to those subjects in the Veterans Administration trial. He may prove not to have hypertension after hospital admission or he may respond to prolonged placebo therapy. Also, the point of muffling of sound may be used to determine a (higher) diastolic pressure in the doctor's office. A patient with a casual diastolic pressure of 100 mm Hg may therefore be equivalent to a patient with diastolic blood pressure of only 90 mm Hg in the Veterans Administration trial. If the doctor remembers to allow for the possi-

Table 5-1. Criteria for withdrawal from the Veterans Administration Co-operative Study Group on Antihypertensive Agents trial during the placebo run-in period.

-
1. Failure to appear for a regularly scheduled clinic appointment
 2. Failure of the urine to contain the prescribed placebo
 3. Failure on a tablet count
 - a. Over 10% too many tablets left
 - b. Five percent or more too few tablets left
-

ble effects of hospitalisation, placebo, and different measurement techniques he must also recall that 58 percent of the Veterans Administration patients had been categorised as having a preceding cardiac, central nervous system, or renal abnormality [57]. The doctor must ask himself how valid are the results of the trial for the patient confronting him in his office.

5.3.2 Subject cooperation and compliance

Those entering a trial tend to be cooperative and willing to make several visits, undergo investigations, and sign consent forms. They are probably more likely to adhere to therapeutic advice than the average patient. The adherence to advice has been termed *compliance* and a trial result may not be valid for a more representative group of patients including a high proportion of noncompliant individuals. In addition, a trial may be specifically designed to exclude noncompliant patients. Table 5-1 gives the withdrawal criteria for the Veterans Administration trial discussed above. Nearly half of the eligible patients were excluded for not attending a clinic appointment, or for not having any placebo marker in their urine, or for failing to produce an approximately correct number of remaining placebo tablets. The trial was a test of active treatment in those who were prepared to take antihypertensive medication regularly and the results are valid for those who do so. This is not a criticism of the trial but there is a limit to how far we can generalise from a particular group of patients to the general population. The results for male veterans are not necessarily valid for women, nor are the results applicable to patients who usually forget to take their tablets. From the results of such a trial, it will be difficult to estimate the response in a group of patients including noncompliant persons. This problem is discussed more fully in section 14.4.

5.3.3 Results are only applicable to certain subgroups of patients in the trial

A difficult problem is provided by a trial with a significant overall result that relies heavily, after subgroup analysis, on some subgroups and not others. For example, in the Hypertension Detection and Follow-up Program trial [58], an overall beneficial result was observed but not for the subgroup of white women. It may be dangerous to generalise from the overall result when such subgroup differences are apparent.

5.4 CONCLUSIONS

The results of a trial may be invalid if the trial is not performed correctly. The provision of adequate controls, avoidance of bias in the results, and reduction in variability are discussed in chapters 7, 8, and 9. In addition, the results of the trial must not be extrapolated beyond the observations made; some examples have been presented in this chapter. The investigator must refrain from jumping to conclusions that he cannot support and must not assume that the results of his trial are valid for subjects of a different age, sex, or race, or noncompliant persons. However, the results of a particular trial can be expected to be repeatable for a demographically similar group of subjects selected and treated in an identical fashion. The results should be valid for such a group of subjects.

6. RECRUITMENT OF SUBJECTS

The recruitment of subjects must be considered very carefully to ensure that sufficient persons with the characteristics required by the investigators are enrolled within an appropriate period of time. The number of subjects is of great importance and will be discussed in chapter 10.

6.1 SUBJECTS WITH THE REQUIRED CHARACTERISTICS

6.1.1 General selection criteria

Selection criteria are discussed in section 3.9. These criteria are required to ensure that the subject has the condition being investigated in the trial and that he cannot be expected to experience an adverse event as a consequence of entering the trial. The general selection criteria will usually include the sex, age, and race of the subjects.

6.1.2 Criteria not usually defined in the protocol

The type of patients recruited will greatly affect the generality of the results (chapter 5). The protocol may not state the social class of those to be recruited and there has been considerable anxiety in the United States that clinical trials are conducted on "people who are least likely to complain or who are least likely to have the power to make their objections felt" [59]. On the other hand, Jeremiah Stamler reported that in the Hypertension Detection and Follow-up Program (HIDFP), "we elected to have a sizable group of (black) patients from the slums of Baltimore, Md., Birmingham, Ala., and Washington, D.C."

[60]. When recruiting patients it may be important to consider social class and other features relevant to generality and the recruitment policy for the HDFP trial allowed the outcome to be compared for whites and blacks.

6.2 THE PERIOD OF RECRUITMENT

The period of recruitment is often not stated in advance and this may have severe disadvantages as the interval tends to become prolonged. Fast recruitment reduces the total length of the trial, the costs involved, and the difficulty of keeping the investigators' enthusiasm at a high level. Even when recruitment times are fixed in advance for a given number of patients, these times can be easily doubled. For example, recruitment to the LRC Coronary Prevention Trial increased from one to two and one-half years [61]. In the Aspirin Myocardial Infarction Study (AMIS) recruitment was limited to one year but was slow at first with a sharp rise prior to the deadline. "Doctors proved to be crisis orientated and most of the recruitment came in the last quarter of the recruitment phase" [62].

6.3 METHODS OF RECRUITMENT

Patients have been recruited for randomised controlled trials from the investigator's own medical practice and from medical colleagues. Suitable patients have also been identified from medical records, from screening programs, and even from volunteers responding to advertisements.

6.3.1 From the investigator's own practice

Patients are most frequently selected from the investigator's own practice. This method may be very effective when only small numbers of patients are required but is rarely applicable to large-scale trials.

The investigator will already be known to the patient and hold a position of trust. This relationship may render the patient anxious to take part in a trial in order to please the investigator and great care has to be taken not to coerce the slightly unwilling patient into taking part.

6.3.2 Referrals from other medical colleagues

The details of the trial must be explained to colleagues who are not taking part in the trial and despite a considerable amount of effort, very few referrals may be forthcoming [61, 62]. Independent clinicians may not be motivated to refer their patients and may forget about the existence of the trial. When practitioners receive fees for items of service, referral of patients to the trial investigators may deprive the referring clinician of financial income for the duration of the trial because many trials provide patients with free attention and medication.

6.3.3 Identification of possible patients from medical records

Physicians' notes and laboratory records have been examined in order to identify suitable patients.

6.3.3.1 Physicians' notes

Physicians' records were examined in the AMIS trial of secondary prevention of myocardial infarction. The notes usually gave the entry criteria of age, sex, and electrocardiographic and enzyme changes so that suitable patients could be selected. Schoenberger [62] reported that this was a better method of recruitment than physician referral and the medical profession was not opposed to this method of selection. However, only 10–20 percent of those patients invited to take part actually entered the trial.

6.3.3.2 Laboratory records

In the LRC Coronary Prevention Trial, patients with type II hyperlipidaemia were entered into the trial. Having failed to achieve many physician referrals, the investigators asked commercial laboratories for suitable patients [61]. The laboratories appeared to divide into three groups.

1. One laboratory wanted payment for the names (this laboratory was not used).
2. There were laboratories who felt a breach of confidence could occur but who agreed to write to the physicians caring for the identified patient to suggest referral for the trial.
3. One laboratory allowed the trial staff to review the results and write directly to the physician.

In this trial both physician and laboratory referrals led to the entry of very few patients.

6.3.4 Screening to detect suitable subjects

Having recruited only 11 patients in Baltimore the LRC Coronary Prevention Trial employed screening methods to search for suitable male patients with elevated serum cholesterol concentrations. These methods are given in table 6–1. Screening of men attending public gatherings, living in Columbia, Md., being considered for another trial or donating blood for the Red Cross led to the recruitment of a further 246 subjects. The screening required a recruitment coordinator, a blood-drawing team, and contact with the news media and film makers for television in order to advertise the activity.

6.3.5 Volunteers with the disease under study (self-referral)

Many patients entered into the AMIS trial referred themselves for inclusion in the trial. Patients were eligible for this trial if they had sustained a myocardial infarct in the preceding five years and volunteers were sought by the advertising methods in table 6–2 [62]. The patients' doctors supported the trial and over a third of those randomised were recruited in this fashion. However, a high proportion of volunteers were not suitable; some of these were excluded when they were first interviewed by telephone.

Table 6-1. Methods of screening men for the LRC Coronary Prevential Trial in Baltimore [61]

A. Screening at:

1. Health fairs
2. Before-church groups
3. YMCA meetings
4. Shopping centres
5. Baseball games
6. 57 Baltimore industries
7. American Heart Association local programs

B. Screening of:

1. The entire city of Columbia, Maryland
2. Blood from American Red Cross donors
3. Potential candidates from the Multiple Risk Factor Intervention Trial (MRFIT)
4. Volunteers responding to mass newsletter mailings

Table 6-2. Methods of getting self-referrals for the AMIS trial [62]

1. *National Media Coverage*

2. *Locally*

- Mass mailings in utility bills
- Public rallies
- Radio and TV announcements
- Newspaper articles
- Paid advertisements

6.4 FACTORS RELATED TO RECRUITMENT

Recruitment can be increased by reducing the exclusion criteria, offering financial support to the investigators, or threatening to withdraw support. The principal investigators also have a great responsibility in generating enthusiasm by visiting, lecturing, and publishing preliminary information.

6.4.1 Changes in protocol

Making the entry criteria less stringent can increase recruitment. For example, in the National Co-operative Gallstone Study recruitment was increased by raising the upper age limit from 69 to 79 [63].

6.4.2 Financial

In the Gallstone Study recruitment increased after it was threatened that research contracts would not be renewed. Similarly, in the AMIS trial contract support was reduced for centres who did not recruit as many patients as they should, and centres who recruited an excess were given increased contract funding.

6.4.3 Threats to withdraw a centre from the trial

In the AMIS trial all centres had to achieve a critical number of patients to remain in the study. It is not known whether increased recruitment in clinics

who were stimulated to achieve the critical number compensated for the number of patients lost in clinics who were removed from the study.

6.5 FACTORS NOT RELATED TO RECRUITMENT

Croke has reported some factors that did not influence recruitment to the various centres in the National Co-operative Gallstone Study [63]: population density, frequency of performing cholecystectomies, expertise of the clinical directors, and incentives provided by ancillary studies were not related to recruitment rates.

6.6 CONCLUSIONS

Recruitment is always a major difficulty in clinical trials. Muench's third law states, "the number of patients promised for a clinical trial must be divided by a factor of at least 10" [64].

In large trials recruitment may occur from clinical practice, laboratory records, population screening, and volunteers with the appropriate disease. Recruitment may be increased by widening the entry criteria and by financial incentives.

Richard Peto, speaking at a meeting in Lyons in November 1981, emphasized the importance of recruiting sufficient patients for trials in cancer patients. He suggested that 1,000 patients are usually the minimum number required and that recruitment can be increased by collaboration between centres, simplification of entry and follow-up procedures, and possibly, by payments to compensate for secretarial expenses.

7. HOW TO ENSURE THAT THE CONTROL AND TREATED PATIENTS ARE SIMILAR IN ALL IMPORTANT RESPECTS

It is essential that the treated and control patients are similar in order that any differences in outcome can be attributed to the treatment and not to other factors. Thus far everyone agrees, but how to obtain similar groups is open to some discussion. In this chapter we discuss the two classical methods: (1) using the patient as his own control (cross-over studies) and (2) random (chance) allocation of patients to distinct and concurrently treated groups. Randomisation is expected to result in the groups being similar. The futility of using historical controls is discussed in section 8.2.3; this chapter deals mainly with the advantages, disadvantages, and methods of randomisation.

7.1 THE PATIENT AS HIS OWN CONTROL

When a patient is given one treatment and then a second the patient acts as his own control. The trial is known as a cross-over trial and the design is discussed in section 11.2. The order in which the treatment is given may be important if the baseline is changing with time (section 11.2.5), and randomisation can be employed to ensure that the patients who receive a certain treatment first are similar to those who initially receive a different treatment. A cross-over design is only feasible for a chronic condition that reverts to its original state with the cessation of treatment (for example, high blood pressure or diabetes mellitus).

7.2 RANDOMISATION

The technique of randomisation in clinical trials refers to the "assignment of treatments to patients using a chance procedure" [65]. Randomisation elimi-

nates bias in patient allocation, renders the treatment groups equal in most important respects, ensures statistical tests are valid, and improves the acceptance of results. It also has disadvantages.

7.2.1 Treatment allocation is free of bias

If an investigator allocates patients to either of two treatments and not at random, he may do so on certain criteria. For example, he may allocate on the basis of the severity of disease. In section 2.2 we discussed the fact that James Lind allocated patients with the most severe scurvy to a particular treatment. The results in such a severely affected group may be biased against Lind's selected and possibly favourite treatment. A more recent example was given by a trial of anticoagulant therapy in myocardial infarction [66] conducted by the Committee on Anticoagulants of the American Heart Association. The investigators considered that there was sufficient doubt as to the efficacy of anticoagulant therapy to warrant a clinical trial of this therapy against the usual basic medical care. Unfortunately, they did not randomise the patients but allocated those admitted on even days to the control group and those admitted on odd days to the anticoagulant group. The results were published and the patients given anticoagulants fared much better than the controls. However, it was noticed that many more patients were allocated to anticoagulant therapy than to control treatment (a ratio of 1.4:1, table 7-1).

It appeared that some referring doctors favoured anticoagulants, and if a patient had a myocardial infarction on an even day they preferred to admit

Table 7-1. The number of episodes of myocardial infarction involved in the American Heart Association trial of anticoagulants [66].

	Excluded	Control group	Treated group
Admitted on even days (total 490)			
1. Not given anticoagulants		395	
2. Given a/c's for complications		35	
3. Given a/c's to prevent complications			31
4. Excluded	29		
Total 490			
Admitted on odd days (total 604)			
1. Given anticoagulants			546
2. Not given a/c's			
Miscellaneous reasons		12	
3. Not given a/c's			
Contraindications			12
4. Excluded	34		
Total	63	442	589

Note: Twelve patients had two admissions and one had three. The number of patients involved in the trial was therefore only 1,080.

them on an odd day. Seriously ill patients could not wait and had to be admitted on an even day giving a smaller control group with severe disease who did badly.

Problems were not limited to recruitment in this trial and difficulties arose both in performance and analysis. Patients in the control group could be given anticoagulants at the request of a private physician and 31 such patients were transferred to the treated group. Moreover, 35 patients who developed thromboembolic disease had to receive anticoagulants but remained in the control group. If the protocol cannot be followed after entry to the trial and the patient is withdrawn or transferred to another treatment group, then the patients should remain in their original randomisation group in order to avoid bias (see the intention-to-treat principle, section 15.7). This rule was followed for 35 patients given anticoagulants in the control group and 12 not given anticoagulants in the treated group owing to contraindications. However, the 31 episodes of active treatment in the control group discussed earlier and 12 episodes of control treatment in the treated group were not analysed on the intention-to-treat principle. Lastly, 12 patients had contraindications to anticoagulant therapy and should not have entered the trial (section 3.9). We should refer to episodes rather than patients when describing the results of this trial as 13 patients were allowed to enter the trial more than once. This duplication should not be allowed (section 15.1.1).

The benefit from treatment reported from this trial was a 27 percent reduction of mortality in men and a 36 percent reduction in women. The greater benefit in women does not agree with other studies and the ratio of treated to controls entered in this trial was 1.5:1 for women and 1.3:1 for men suggesting that the failure to randomise may have distorted the results more for women than for men.

7.2.2 The treatment groups are similar with respect to important confounding variables

Confounding variables are those factors other than treatment that are known to be related to the outcome. They are otherwise known as interfering or nuisance variables and the effect of such variables is to confound (alter) the effect of treatment. If, for example, survival from cardiovascular disease is being examined according to whether or not the patients are given a lipid-lowering drug or placebo, it is important that the active and placebo groups are similar for the known risk factors for cardiovascular disease such as high blood pressure or cigarette smoking. Randomisation can be expected to lead to two groups similar with respect to these characteristics.

If the groups are similar, it does not imply that they are identical. However, it can be anticipated that the two groups, for a given confounding variable, will not differ at the 5 percent level of significance with respect to this item. When considering a large number of confounding variables it is unlikely that randomisation will ensure that the groups are similar in every respect; in fact,

with 14 variables there is a greater than even chance of one item differing significantly at the 5 percent level. But most confounding variables will not differ with statistical significance between the two groups. On the other hand, nonsignificant differences may be of biological importance. This is discussed in section 7.2.5.

7.2.3 Randomisation ensures that many statistical tests are valid

Many statistical tests assume that the populations to be compared are randomly drawn from a larger single population. This is manifestly true for randomised controlled clinical trials where the treatment groups are randomly drawn from the total group of patients.

7.2.4 Randomisation improves the chance that the results will be accepted

Weinstein [67] has pointed out “one randomised trial with 100 patients can dramatically change physician behaviour, whereas the experience of 100,000 patients might be neglected.” Aware that many accept new treatment uncritically, the physician will naturally be influenced more by an objective unbiased study such as the randomised controlled trial. When randomisation has not been carried out but the data adjusted afterwards to correct for differences between the groups, Weinstein went on to say, “Although matching or covariance analysis may satisfy the investigator himself that all important nuisance variables have been washed out of the analysis, there will always be someone somewhere who will complain that the analysis did not control for colour of eyes or sunspots . . . that he believes . . . to be confounding the results.” Randomisation usually provides groups similar with respect to the colour of their eyes—and sunspots!

7.2.5 Disadvantages of randomisation

7.2.5.1 Randomisation may lead to unequal numbers in each group

When large numbers of patients have to be randomised, the numbers in each group are likely to be very similar. With small numbers (as in a small trial or in one centre of a multicentre trial), the number receiving one treatment may differ considerably from those receiving a second treatment. Zelen [65] gives an example for 24 patients where random allocation led to nine on one treatment and 15 on another. This problem may be overcome by restricted randomisation (section 7.3.4).

7.2.5.2 Randomisation may still provide groups that differ in some important respect

After randomisation, bad luck may lead to the groups being different for an important variable and such a difference may be both statistically and biologically important. With small numbers the differences may not be statistically significant, but they may still be of biological importance. Considering the previous example of 24 patients, it is possible that even if 12 patients received one

and 12 the other drug, eight of one group but only four of the other group could be male. This difference would not reach the five percent level of statistical significance but the two groups could hardly be said to be similar with respect to the proportion of men in each group.

Important confounding variables can be allowed for prospectively by stratification (that is, randomisation within certain strata such as age groups) (section 7.3.6) or matching one patient with another. Retrospectively an imbalance can be allowed for in the analysis [5].

7.2.5.3 Randomisation alone may be less efficient than matching

Weinstein [67] stated that "If a confounding factor is known to be present, a smaller sample size would be sufficient to achieve comparable statistical information content by matching or adjustment procedures than by randomisation alone." But attempts to get matched patient pairs may have two undesirable consequences. First, as the trial proceeds it may become apparent that another important confounding variable should be considered and the pairs will not be matched for this variable; and second, partners may not be found for certain patients and they may have to be omitted from the trial. Randomisation within a limited number of strata may provide similar groups when more than a few subjects are to be entered.

7.2.5.4 Is it desirable for some patients to be allocated to an inferior treatment by chance?

The result of the trial may show that one treatment is inferior and it has been suggested it is immoral to decide this outcome by chance. Randomisation can be adapted to limit the number of patients who receive an inferior treatment (section 11.8). But if two treatments are not equal some patients must suffer. A superficially attractive way around this problem is to ask the patients to select their treatment. This is usually not appropriate for drug treatment but has been suggested for operative treatment [67]. For example, one operation may be thought to have a high initial mortality but an increased likelihood of long-term survival and a second operation may have a lower initial mortality but worse long-term survival. A young patient with a family may well opt for the second operation. As in this example, it is very unlikely that the patients who opt for one operation will be similar to those who choose the other operation, leading to the problem that the effect of treatment will be confounded (confused) with the different characteristics of the patients. Randomisation remains the only safe method of selecting the two groups and when the investigator is in genuine doubt concerning the preferred treatment, only chance allocation can be ethical.

7.3 METHOD OF RANDOM ALLOCATION

We must discuss at what stage of the trial we should randomise, how to use random number tables for randomisation, whether to restrict randomisation to give an equal allocation of subjects between treatment groups, whether to

choose an unequal allocation, and lastly whether or not to randomise within certain groupings or strata.

7.3.1 The stage of the trial at which a patient should be randomised

In clinical trials the investigator should first decide that a patient is eligible for the trial, then randomly allocate the patient to a treatment group. He must not be aware of the next treatment to be allocated as with this knowledge, bias in allocation may occur. For example, if the investigator is aware that the next patient to be entered will receive placebo treatment, he may be unwilling to enter a patient with severe disease. The placebo group will then contain an excess of less severely affected patients. The investigator must be certain that the patient is suitable for any of the trial treatments and then determine the result of randomisation.

7.3.2 Method of determining the result of randomisation

To ensure that the investigator is not aware of the next treatment to be allocated he may be required to contact a central coordinating office to determine the result of randomisation. Alternatively, if the result of randomisation is held in the investigator's office, the treatment to be allocated should be held in sealed numbered envelopes, opaque when held up to the light, to be opened sequentially as required and only after the patient's name and other details have been written on the appropriate envelope.

7.3.3 Method of simple randomisation

When there are only two treatments, randomisation can be achieved by simply tossing a coin. However, it is usually administratively more simple, and less open to manipulation, to decide the randomisation in advance. This can be done conveniently using a table of random numbers as illustrated by table 7-2, the four-digit random numbers having been generated by a computer program. The table may be read as four-digit numbers or single figures or any number of digits either horizontally or vertically. Table 7-3 gives the results of using single numbers horizontally and allocating treatment A when the random number is even and treatment B when the number is odd. The starting number was selected at random in the first row, the sixth number from the left. After 20 patients the ratio of patients on A:B was 1:1.5 and after 100 patients, 1:1. With a small trial grossly unequal numbers of patients may be allocated to two treatments. This will not be a problem with large trials unless they are multicentre, where the small numbers allocated to individual centres may result in some centres having a markedly unequal distribution of treatments. The problem of unequal treatment groups is overcome by restricted randomisation [68], otherwise known as block randomisation [65].

Table 7-2. Four-digit, computer-generated random numbers

0011	0858	4371	7901	2691
1893	2018	1425	6942	0755
6658	9411	5940	5568	7667
5147	6128	9689	6802	0238
7886	1512	3488	1699	7461
9675	5715	8699	2543	0769
7756	1808	5655	4214	6091
6931	9953	6577	3369	2761
7134	5116	3817	6314	3651
3921	4119	9007	5648	3664
4893	7074	0631	7365	2477
5848	8690	1833	7555	6690
1849	8756	6811	0913	8525
9462	9739	0522	9183	8797
9843	8360	7114	3640	3097
2322	4221	6663	9846	8555
1705	7965	7440	8084	6412
0987	4087	6593	4548	1165
7453	9080	1084	1354	1610

7.3.4 Restricted or block randomisation

Restricted randomisation ensures that within a block of patients equal numbers are allocated to each treatment. For example, when it is decided to restrict the randomisation so that every group of four patients includes two on treatment A and two on B, then Zelen [65] suggested allocating a number to each of the six ways that the treatments can be allocated in groups of four. As before, the random number table is consulted, not to allocate a single patient to a single treatment, but to allocate a sequence of four patients to four treatments as in table 7-4.

The disadvantage of restricted randomisation is that the investigator may be able to predict the next treatment to be allocated. For example, an investigator who is aware that randomisation is restricted for blocks of four subjects would be able to predict the allocation for the fourth, eighth, or twelfth patient, and so on. With a multicentre study, restricted randomisation at the coordinating centre may lead to balanced numbers in the trial as a whole but not for an individual participating centre and the local investigator would be unable to predict the treatment. However, it is desirable to have roughly equal numbers at each centre as patients may do particularly well, or badly, at one centre. If a certain centre has a predominance of one treatment the result may appear to be due to that particular treatment. It is therefore necessary to randomise within each centre to ensure that a roughly balanced allocation occurs at the centres. Zelen has described a method, called balanced block randomisation [65], which gives a balanced allocation for each centre while making it difficult to predict the treatment sequence. The method requires the use of an auxiliary random number table whereas variable block allocation can achieve the same

Table 7-3. Allocation of two treatments, A and B, to 100 patients according to whether the random numbers in table 7-2 are even (A allocated) or odd (B allocated).

Patient	Random Number	Treatment A or B	Cumulative number on	
			A	B
1	8	A	1	0
2	5	B	1	1
3	8	A	2	1
4	4	A	3	1
5	3	B	3	2
6	7	B	3	3
7	1	B	3	4
8	7	B	3	5
9	9	B	3	6
10	0	A	4	6
11	1	B	4	7
12	2	A	5	7
13	6	A	6	7
14	9	B	6	8
15	1	B	6	9
16	1	B	6	10
17	8	A	7	10
18	9	B	7	11
19	3	B	7	12
20	2	A	8	12
...
40	9	B	19	21
...
60	6	A	28	32
...
80	1	B	42	38
...
100	5	B	50	50

Note: Single numbers have been selected working horizontally and starting at the first line (see text).

ends and be more easily understood. Variable block allocation is therefore described here.

7.3.5 Variable block randomisation

In variable block randomisation the investigator does not know the number of patients to be recruited before balance is achieved. Equal numbers may be

Table 7-4. The use of a random number table to achieve restricted randomisation

Random number	Sequence
1	A B A B
2	A A B B
3	B B A A
4	B A A B
5	A B B A
6	B A B A

Note: Randomisation is restricted to give two patients on drug A and two on drug B for every four patients entered into the trial. The random number is selected from a table of random numbers and gives the treatment allocation for the next four patients to be entered into the trial.

Table 7-5. The 20 different possible sequences for two treatments and blocks of six patients.

Allocation number	Treatment sequence
07	A A A B B B
08	A A B A B B
09	A A B B A B
10	A A B B B A
11	A B A A B B
12	A B A B A B
13	A B A B B A
14	A B B A A B
15	A B B B A A
16	A B B A B A
17	B B B A A A
18	B B A B A A
19	B B A A B A
20	B B A A A B
21	B A B B A A
22	B A B A B A
23	B A B A A B
24	B A A B B A
25	B A A A B B
26	B A A B A B

Note: The random number can be searched for in random number tables to decide on the treatment allocation. Random numbers 1-6 can be used to allocate treatment to blocks of four subjects (table 7-4).

reached after, say, four or six patients and this makes it difficult for the investigator to predict the next treatment to be allocated. Balance after four to six subjects can be achieved by extending table 7-4 so that random numbers seven to 26 cover the 20 different sequences for blocks of six (table 7-5). The random numbers in table 7-2 must be examined in pairs and four block or six block sequences allocated according to numbers 01, 02, . . . 26. A further alternative

to variable block randomisation is adaptive randomisation. This is a complex procedure whereby the probability of selecting a treatment can be reduced if an excess of that treatment has been allocated [65].

7.3.6 Randomisation within strata

In the same way as restricted randomisation prevents a disproportionate number of patients being allocated to one treatment, stratification ensures that one treatment group does not include an excess or deficit of subjects with a certain confounding characteristic (for example, male). Inequality between groups is usually only a problem when the numbers in a trial are small. However, most investigators would stratify: (1) by centre (in a multicentre trial); and (2) by sex.

Randomisation is carried out within each centre and for both sexes separately. Weinstein [67] has argued that "matching, blocking or adjusting may be far more efficient devices than purely randomizing. Why let chance do what one can do for oneself?" Matching can be considered to be the most extreme form of stratification with the subjects in matched pairs being randomly allocated, one to receive one treatment and one the other. Retrospective matching or adjusting the data is less desirable than starting with similar groups.

Peto and associates [69, 70] considered stratified allocation to be an unnecessary complication for large trials. In such trials chance will usually ensure that the treatment groups are comparable and retrospective stratification can be used to compare the treatment effects in one stratum (for example, males) with the effect in a second stratum (in this example, females). Stratification or matching cannot ensure that the treatment groups are equal in all important respects. Often the variables that are known to be of importance are too numerous and stratification for these features would lead to several strata containing too few patients. Also, if previously unrecognised but important features are discovered during the course of the trial or during analysis, stratification or matching will not have coped with these. In conclusion, the number of strata should be kept at a minimum and restricted randomisation may be carried out within each strata if desired.

7.3.7 Randomisation schedules that do not allocate two treatments to an equal number of patients

When comparing a new treatment with established therapy, Peto [69] has suggested that the numbers allocated to a new treatment may be increased by giving two patients the new treatment for every one patient given the control treatment. He states, "there will be an unbiased randomised comparison which is very nearly as sensitive as an ordinary equal-groups randomised trial . . ." In long-term trials of survival, when the new treatment is expected to reduce mortality by 50 percent, then 2:1 randomisation may be expected to yield an equal number of deaths in two groups. Two-to-one randomisation can be achieved by using twice as many random numbers for allocation to the

new treatment as are used to allocate to the control treatment. Unequal randomisation is very attractive when little is known about the new treatment whereas a good deal is known about the old or control treatment.

7.4 ALTERNATIVES TO RANDOMISATION

There is usually no good alternative to randomisation. The following section gives some alternatives that have been suggested and the problems that may be encountered with them.

7.4.1 Allocation according to date of entry to the trial

This was used in the trial discussed in section 7.2.1 [66]. The allocation was known in advance, odd dates for one treatment, even dates for another, and the trial entry was manipulated to give unequal groups. Prior random allocation, using random number tables, is simple and not open to such manipulation.

7.4.2 Allocation according to the hospital number

The hospital number has been employed, even numbers being allocated to one treatment, odd numbers to the other. This can be open to manipulation as with the date of entry to the trial.

7.4.3 Allocation according to the initial letter of the subject's name

If the initial letter of the subject's name is A to M, he can be allocated to one treatment, and if it is N to Z he will go to the other. This is a very unsatisfactory method. Not only can allocation be manipulated but the treatment groups may be very different. For example, in the United Kingdom there are more patients with surnames starting A to M than N to Z. Moreover, A to M will include an excess of Scotsmen with names starting with Mc; N to Z will include a greater proportion of Irishmen with names starting with O' and also Asians with the surname Singh or Patel.

7.4.4 Allocation according to the wishes of the patient

This procedure has been proposed in the context of trials of surgical procedures [67]. Although it is desirable to conform with the patient's wishes, it is both possible and probable that patients choosing one form of treatment will differ from those choosing another. Obvious differences can be adjusted for retrospectively in the analysis, but differences may remain that are not recognised and can be confused with treatment effects. Random allocation is by far the safest procedure and cannot be recommended too strongly.

7.4.5 Allocation according to the preceding results

An example of such a strategy is given by the play-the-winner rule: if a treatment is followed by success the next patient also receives this treatment; if a treatment results in failure the next patient receives the alternative treatment.

Play-the-winner rules are discussed in section 11.8. Not only must the results be known quickly but the allocation may be open to manipulation.

7.5 CONCLUSIONS

It cannot be stressed too strongly that randomisation is necessary to ensure that the different treatment groups are similar in most respects. Occasionally bad luck will still lead to unequal groups and retrospective adjustment of the data becomes necessary. However, randomisation is the best safeguard that the groups are equal and that the investigator does not consciously or unconsciously manipulate entry to the trial, thereby producing groups of unequal numbers or differing confounding factors.

A subject must be eligible for the trial and *then* randomised to a treatment group. Randomisation may be restricted to give equal numbers in each treatment group and restricted randomisation may be employed within strata to ensure equality of confounding variables. Randomisation can be employed when there are more than two treatment groups and when more patients are to be allocated to one treatment group than another. In a within-patient cross-over trial, randomisation is also required to determine the order of treatment and to make sure that whether treatment A precedes treatment B, or vice versa, is a matter of chance and not open to manipulation by the investigator.

8. HOW TO ENSURE THAT THE RESULTS ARE FREE OF BIAS

Biased results are distorted and prejudiced in favour of one treatment or another. When discussing the assessment of a patient's progress in a trial, Bradford Hill [8] wrote, "the judgements must be made without any possibility of bias, without any overcompensation for a possible bias, and without any possibility of accusation of bias."

The results of a trial may be biased if patient allocation favours one group rather than another; if an adequate control group is not provided and therefore the results cannot be interpreted; when noncompliance influences the results; when patient or investigator bias affects the assessments; and when analytical bias distorts the presentation of the data.

8.1 BIAS DUE TO AN UNEVEN ALLOCATION TO THE TREATMENT GROUPS

This bias may arise due to the failure or absence of randomisation and is dealt with in chapter 7. To summarise, the patient must first be considered eligible for the trial and then randomised to his particular treatment group. The investigator must not have prior knowledge of the treatment that will result from randomisation.

8.2 AN ADEQUATE CONTROL GROUP MUST BE PROVIDED

Controls are necessary in clinical trials as otherwise any improvement or deterioration associated with giving the experimental treatment cannot be confidently attributed to the treatment. Randomised controlled trials rarely

show as much benefit from treatment as most investigators assume will be the case. There are several reasons why an experimental treatment appears so much more effective in observational studies than in a controlled trial and these include the placebo effect, regression to the mean, and time trends in the condition being studied.

8.2.1 The placebo effect

In section 2.4 two early examples of placebo responses were reported; Haygarth's use of dummy wooden appliances to investigate the use of Perkin's metal tractors which were supposed to cure by electricity [14] and Sutton's report of the effect of mint water in rheumatic fever [15].

In the 1920s the Western Electric Company carried out some experiments in its Hawthorne plant in Chicago. Illumination was either increased, decreased, or held constant. The workers were interviewed and the increased attention paid to them led to a rise in production, independent of changes in lighting intensity [71]. There was therefore an unplanned effect in the control group and this became known as the Hawthorne effect. Ederer [68] considered that the placebo effect may arise either from a change in the social situation or from suggestion. Table 8-1 gives measurements of placebo effects mostly reported in a review by Beecher [72]. Even severe postoperative pain and angina can be satisfactorily relieved by placebo in up to 40 percent of patients, headache alleviated in 52 percent, and cough in 37-40 percent.

Shapiro [85] has concluded, "we are led to the conclusion that the history of

Table 8-1. Placebo response in painful conditions and cough (largely derived from a table by Beecher [72]).

Complaint	Placebo	Percentage patients responding to placebo	Reference
Postoperative pain	Intravenous	21	73
	Saline	26	74
	Subcutaneous	31	75
	Saline	39	76
	Lactose by mouth	33	77
Angina pain	Tablets by mouth	26	78
		38	79
		38	80
Myocardial infarction	Intravenous saline	54(10 min) 31(10-30 min)	81
Headache	Lactose by mouth	52	82
Cough	Lactose by mouth	40	83
	Subcutaneous saline	37	84

medical treatment can be characterized as the history of the placebo effect, since almost all medications until relatively recently were placebo." Dollery [86] suggested that the placebo effect can be both psychological and physical. He considered that "A friendly smiling face, comforting personality and a caring environment make most people feel better. The effect is reinforced if the caring environment is combined with the powerful suggestion that a specific treatment will be of benefit." However, Dollery proceeded to point out that physical changes can result from suggestion. He commented that hypnosis can lower airways resistance in some patients with asthma and that a placebo may stimulate the release of encephalins in the brain and thereby have a true pain alleviating effect. Enkephalins act on opiate receptors in the brain and when naloxone, a drug that blocks these receptors, was compared with placebo to assess the pain relieving effect of this drug, it was found that placebo was more effective. If a placebo can be so powerful, it is imperative to employ a placebo or other simultaneously treated control group so that the effect of the experimental medicine can be shown to be greater than the control treatment. The method of employing placebo treatment is discussed in section 8.10.

8.2.2 Regression to the mean

In a medical practice, patients will usually be seen when they have developed symptoms. If the patients are suffering from a self-limiting illness, then they can be expected to improve and any treatment prescribed may be given the credit for this improvement. Similarly a patient may have a chronic condition that varies in severity so that he will see the doctor when the disease is most troublesome and then be expected to improve. Again, treatment may receive the credit. The phenomenon of selecting a group of patients when they have a high recording of some measurement (for example, number of symptoms or blood pressure) and then observing an overall improvement is one form of regression to the mean (sections 9.1.4 and 14.2.3).

Blood pressure varies considerably and if a group is selected when the measurement is high, it can be expected that next time the average pressure will be lower. Similarly if a group is selected because the reading is low, the subsequent mean reading will be higher. Regression to the mean is the movement of a result from an outlying position nearer to a more moderate value as a consequence of an initial selection of the subjects.

8.2.3 Time trends

When historical controls are employed, a reduction in the severity of disease with time may bias the interpretation of results in a group being studied later. This problem has been reviewed by Peto and his colleagues [70] who concluded that the latter group may also be affected by changing referral patterns, changes in ancillary treatment, and the selection of patients. They point out that ignoring a few old or untreatable patients in the new series may make a

large difference. The comment of the group, however Peto and co-workers considered that historical controls are not useful in assessing the results of randomised trials. For example, if randomised controls have fared worse or better than expected it may cast doubt on the trial conclusions. The control group in the University Group Diabetes Program trial did better than expected but this criticism has not been very persuasive (section 19.6). On the other hand, the fallacy of using a historical control as the only control has been conclusively demonstrated [5, 70, 87, 88].

8.3 NONCOMPLIANCE MAY BIAS THE RESULTS

Noncompliance may be defined as nonadherence to therapeutic advice and its occurrence may radically alter the interpretation of a trial result. Figure 8-1 illustrates two possible assessments of the effect of stopping smoking on total mortality in men; the first arose from an observational study of the difference between subjects who stopped smoking and those who continued and the second assessment resulted from a randomised controlled trial. The observational study [89] reported an excess of mortality in those who continued to smoke compared with those who spontaneously stopped smoking (0.15 percent/year). The observational study suggested a benefit from stopping smoking whereas the randomised controlled trial revealed a slight and nonsignificant increase in total mortality in those who were given antismoking advice [56]. The result of the trial did not fulfill the hopes raised by the observational study. A longer period of observation may reveal different results, especially as any adverse effects of smoking, if reversible, may take time to be corrected.

A further reason for the poor result may have been noncompliance with the therapeutic advice. Not all the intervention group stopped smoking (only 36 percent had stopped three years after the advice was given). Also some of the control group did stop smoking (14 percent after three years). However, these facts may not explain all the differences between the results of the observational and intervention studies. In an intervention study subjects may be persuaded to stop smoking who would not otherwise do so. In the observational study, those who stopped smoking were presumably a health-conscious group who may also have been moderate in their diets and taken exercise. Not surprisingly these men had a low mortality and it must be noted that subjects who never smoked had the lowest mortality. The randomised controlled trial was the best test of the effect of intervention and tested the strategy of giving antismoking advice. The strategy had only a modest success, the majority of subjects being noncompliant with the therapeutic advice. In this example, noncompliance can be expected to result in a smaller effect of intervention than predicted when observing only subjects who stopped smoking.

It would appear attractive to assess the effect of smoking only in those who take the advice; this is equivalent to considering only those who are compliant with treatment. This analysis would be appropriate if the trial tests only for a relationship between stopping smoking and lower mortality, an explanatory

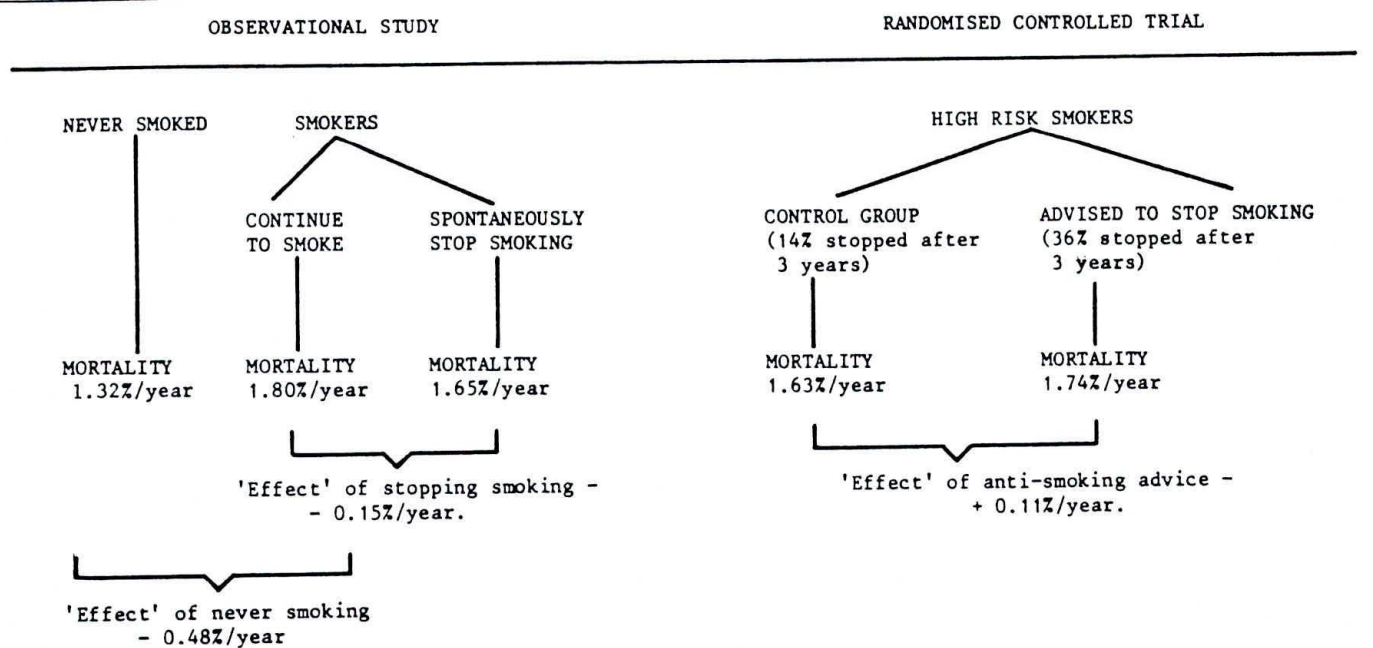


Figure 8-1. The effects of stopping smoking as determined from an observational study and a randomised controlled trial (see text).

trial (section 10.5.1). However, the experimental treatment in the trial was giving antismoking advice, expecting to demonstrate a causal association that could be reversed. If only those who stopped smoking were analysed, the results would apply to men who were relatively health-conscious: the same group that was assessed in the observational study.

The results of a randomised controlled trial may be worse than expected when noncompliance is a problem or better than expected when compliance is improved under trial conditions. The results of a trial must be analysed primarily on the intention-to-treat principle (section 15.7) in order to avoid bias from selecting a group with an increased compliance.

8.4 PATIENT BIAS MUST NOT INFLUENCE THE RESULTS

In section 8.2 we discussed the strength of the placebo response. During treatment the patients wish to feel better and will tend to report favourably, implying an improvement in their clinical condition and also pleasing the investigator to whom they may feel indebted. Ederer [68] quoted Francis Bacon: "For what a man had rather were true he more readily believes." Shapiro [85] stated, "The desire for successful outcome is felt so strongly in both patient and investigator that objectivity cannot be guaranteed. Both have an emotional stake, overt or occult, in the result. Further, the giving of any treatment especially by needle injection, is a strong psychotherapeutic stimulus in itself."

An adequate control group has to be employed to measure any placebo effect. However, the placebo effect will almost certainly not be observed if the patient knows that a particular treatment is a dummy. Therefore the patient must not know which treatment he is receiving and he is said to be blind with respect to the treatment. When the patient is not aware of the nature of his treatment but the investigator does know, a trial is said to be single-blind. When both the patient and the investigator are unaware as to the treatment being given, the trial is said to be double-blind. Investigators in ophthalmological research prefer the term *doubly masked* owing to the disagreeable association and possible confusion that may result from the use of the term *blind* [90]. The term *treble blindness* has also been advocated and can be used when a statistician analyses the results before being aware of which treatment was allocated to the groups in an otherwise double-blind trial.

8.5 INVESTIGATOR BIAS MUST NOT INFLUENCE THE RESULTS

Few would argue that the patient is an unbiased observer of his well-being, the placebo effect being too well recognised. However, it may be worth discussing how difficult it may be for an investigator to make unbiased observations. Ederer [68] discussed the discovery of n-rays by the eminent French physicist Blondlot. These rays were said to be given off by metals and to increase the luminosity of cards painted with luminous paint. In addition, the rays, when falling on the eye, increased the person's ability to see objects in a nearly dark

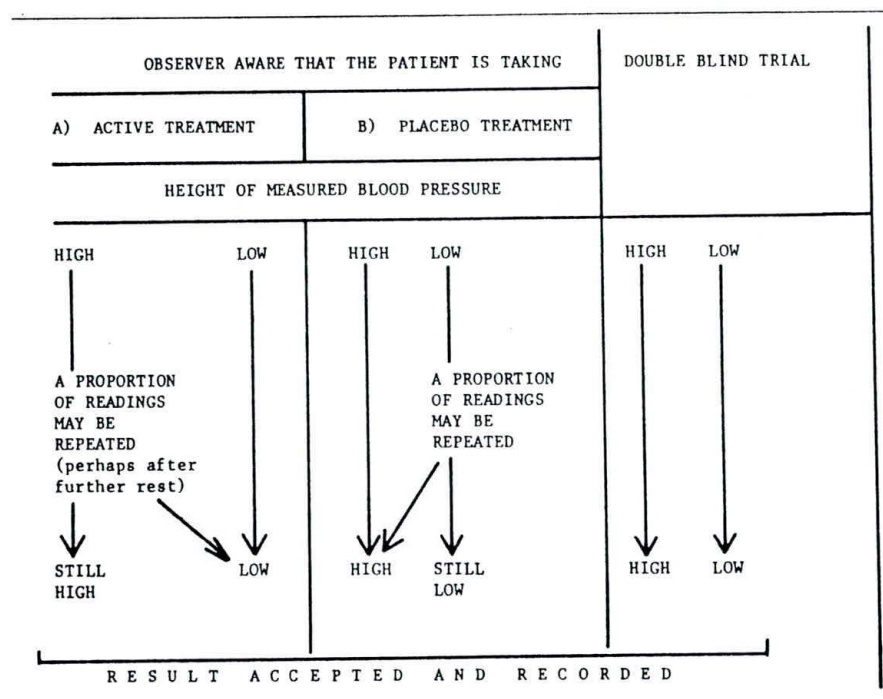


Figure 8-2. Hypothetical blood-pressure readings in an open and a double-blind trial.

room. The existence of these rays was disproved by a double-blind study by Pozdena [91] and also by a trick in which the American physicist, Wood, substituted a wooden ruler for a metal file and asked Blondlot whether the file was producing n-rays. Blondlot assured Wood that the file enabled him to see much better [92]. Pozdena performed a straightforward but elegant double-blind trial in which an assistant opened and closed a shutter, releasing and interrupting the flow of hypothetical n-rays. Pozdena recorded increased luminosity as often when the shutter was open as when it was closed [91].

Blondlot's error arose from the difficulty of making subjective assessments. Observer bias can present great problems when determining subjective impressions such as the presence of symptom side effects; this is discussed in chapter 16. More objective measures may also be influenced by observer bias; section 9.1 discusses the repeatability of measurements, and section 9.2 the quality control of data. The measurement of blood pressure in a trial of an antihypertensive drug may be taken to illustrate the difficulties. Let us assume that the protocol specifies that the blood pressure has to be taken after 5 minutes rest in the lying position. Figure 8-2 illustrates the possible sequences of events, first when the observer is aware of the treatment being prescribed and second when the trial is double-blind and the observer is not aware of the

treatment being given. For each instance the hypothetical result of detecting a high or low reading is charted. It is possible that all first readings are accepted and recorded except for high readings when the patient is known to be receiving active treatment. In this case the observer may assume, possibly rightly, that the patient has not relaxed sufficiently or that the blood pressure has not been taken correctly. In this case the cuff may be applied more carefully (for example, with the inflatable section more accurately over the brachial artery); the patient may be asked to relax for a further two minutes, and the readings repeated. We assume that the second reading is accepted and recorded. The observer may be correct when he substitutes the second lower reading for the initial high reading. However, high measurements will not necessarily be repeated when the patient is known to be taking a placebo and the blood pressure recordings, on average, will be biased towards lower readings on the active treatment. Similarly, it is possible that low readings on placebo will be repeated, making matters even worse. No such bias is to be expected in the double-blind trial.

Fletcher [93] stated, "Both in initial assessment of the patients and the subsequent assessment of their progress the tests should be applied by observers who remain unaware of which patient is undergoing treatment and which is a control. If this is not done, the subjective judgements which are inseparable from nearly all tests in clinical medicine may prejudice the results . . ."

Another example where objective measurements are subject to observer bias was given by Kahn and his colleagues [94] when reporting on measurements of serum cholesterol. They found that when technicians were given blind duplicate bloods to measure, the standard deviation of the duplicates was 2.5 times as large as when they were given labelled duplicates. Wilson concluded, "No human being is ever approximately free from these subjective influences; the honest and enlightened investigator devises the experiment so that his own prejudices cannot influence the result. Only the naive or dishonest claim that their own objectivity is a sufficient safeguard . . ." [95]. Observer bias can favour or prejudice a positive result.

8.5.1 Observer bias favouring a positive result

Muench's Second Law states: "Results can always be improved by omitting controls" [96]. A control group can demonstrate the occurrence of spontaneous improvement and thereby reduce the magnitude of any overestimated treatment effect. Even with controls, the observer may tend to report more favourably on a new treatment under investigation. When the trial is double-blind, observer bias may be prevented and produce a more correct estimate of the effect of treatment.

Foulds reviewed studies of antidepressant drugs conducted between the years 1951 and 1956 [97]. He identified 36 studies in the American literature and 36 in British journals. Only four trials in the American literature included controls. In the British literature 16 trials included controls and, on average,

these papers reported a 19 percent success from treatment. In contrast, the 20 uncontrolled studies reported an 85 percent success rate. Foulds also made the interesting comment that patients with anxiety states tended to improve spontaneously, thus when subgroups were analysed these patients appeared to have responded best to treatment. Without adequate controls, it might be erroneously concluded that the antidepressant drugs were most effective in patients with anxiety states.

8.5.2 Observer bias prejudicing a positive result

This theoretical consideration has been discussed by Ederer [68], who suggested that a bias may occur against the experimental treatment "when the investigator's bias is against the treatment or when he overcompensates for his known bias in favour of treatment. Thurber's moral, 'you might as well fall flat on your face as lean over far backward' [is apt]."

8.6 BIAS ARISING FROM THE ANALYSIS

There are many ways in which the analysis of trial results could bias the conclusions. An example was discussed in section 8.3. In the trial of antismoking advice [56] the results in the intervention group could be improved by omitting those who did not stop smoking and made worse in the control group by leaving out those who did stop smoking. The intervention, however, was the giving of antismoking advice and the total intervention group has to be compared with the total control group as the trial was not intended to be an explanatory study. Similarly, the Anturane Reinfarction Trial [48] has been criticised for leaving out patients in the intervention group simply because they died while the serum concentration of Anturane (sulfapyrazone) was thought to be negligible (see section 19.1). It has therefore been suggested that trials should be analysed without the statistician's being aware of the treatment given to the various groups, the treble blindness defined in section 8.4.

The greatest problem in analysis arises from the difficulty in deciding whom to include in the analysis and whom to omit. Similarly during the course of the trial withdrawal of patients may bias the results.

8.7 BIAS DUE TO THE WITHDRAWAL OF PATIENTS FROM A TRIAL

Bias may result from withdrawal from the group as a whole (for example, during a placebo run-in period) or by selective withdrawal from the different treatment groups.

8.7.1 Withdrawal of patients during a placebo run-in period

The withdrawal of noncompliant patients prior to randomisation results in compliant subjects being entered into the trial and the conclusions from the trial results are only valid for such patients. The problem is discussed in chapter 5.

8.7.2 Selective withdrawal of patients from the placebo group

Placebo treatment may be less effective than active treatment and patients may be withdrawn owing to treatment failure. In long-term trials of antihypertensive agents to determine their effect on mortality, more patients are withdrawn from the placebo group than from the actively treated group owing to marked rises in blood pressure. If a withdrawal criterion states that a patient must be withdrawn with a certain elevation in blood pressure then the assumption is that, left in the trial, this patient would be likely to suffer an end point such as a stroke. If the withdrawn patients are omitted from the analyses, the placebo and actively treated groups will no longer be comparable for important characteristics, as less patients with increasing blood pressure will be removed from the actively treated group [98]. If withdrawn patients are included in the analysis, it must be assumed that they suffered an adverse event with a defined probability. It would appear unreasonable to define the probability as 1 since no adverse event was in fact observed. An estimate may be available of the probability from observational studies or from earlier trials when patients with the given level of blood pressure were not withdrawn but continued in the study. It may then be possible to allocate some estimate, say, three cardiovascular end points for every ten such patients withdrawn from the trial.

8.7.3 Selective withdrawal of patients from the actively treated group

In a double-blind trial, an excessive withdrawal of patients from the actively treated group may be expected to represent the frequency with which the active treatment produces adverse effects. However, in a trial that is not double-blind patients with possible adverse reactions will only be withdrawn from the actively treated group.

In a single-blind trial a patient may suffer an episode of illness that he attributes to the drug he is receiving. Figure 8-3 illustrated the sequence of events that may occur. The patient may have a gastrointestinal upset, rash, influenza, or other condition unrelated to treatment. If he has just started therapy, he may attribute the symptom to the drug or other treatment and not to an intercurrent illness. When the patient is on active treatment, the doctor may agree with the diagnosis and withdraw the patient from the trial with a possible adverse event. When on placebo treatment, the patient will not be withdrawn and the losses from active treatment will be greatly in excess of the withdrawals from placebo treatment. This situation is avoided in double-blind trials.

It must be remembered that placebos do occasionally produce an adverse effect. If a placebo tablet contains lactose it is possible that diarrhoea will be produced in susceptible individuals and in this event any excess incidence of diarrhoea with the active treatment may be underestimated rather than overestimated. Usually however, a placebo group and double-blinding gives an unbiased estimate of the frequency of adverse drug reactions (see chapter 18).

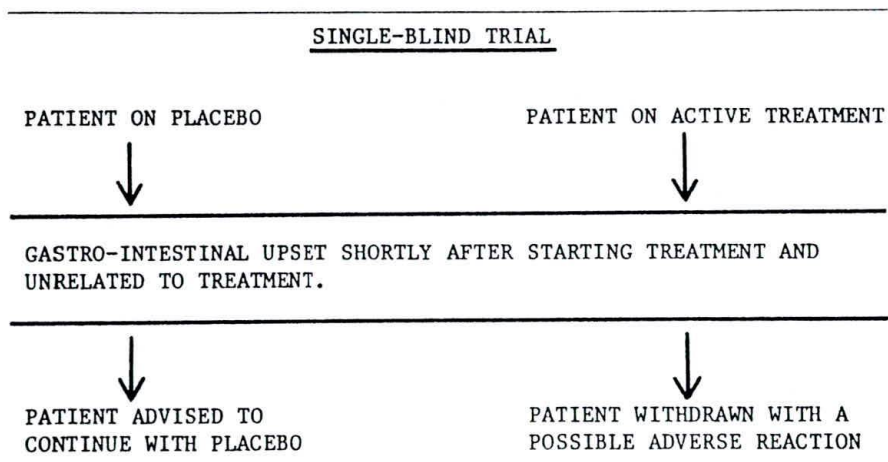


Figure 8-3. Mechanism whereby the adverse effects of treatment may be overestimated in a single-blind trial.

8.8 BIAS DUE TO FAULTY METHODOLOGY

When the trial protocol is not identical for both the treatment under investigation and the control treatment, then biased results may be obtained. Two examples may be provided: (1) when the treatment has an additional but nonspecific effect this may be difficult to reproduce in the control group, and (2) when the protocols for the two groups differ in order to save time or money.

8.8.1 The treatment may have a nonspecific effect that is difficult to reproduce in the control group

The nonspecific improvement resulting from giving drug treatment can be estimated from a placebo group in a double-blind trial of drug treatment. The nonspecific effect cannot be estimated when surgery is the intervention strategy and may also be difficult to assess when therapeutic advice is given. In a trial of antismoking or dietary advice, the patients in the intervention group may be seen repeatedly and it will be difficult to provide placebo counselling on any neutral topic. First, it would not be ethical to advise the control group against an activity that does not harm them. For example, in a controlled trial of antisalt dietary advice in hypertensive patients poorly controlled on antihypertensive drugs, we decided that only the patients randomised to the diet would get dietary counselling. However, interviews were arranged for both the intervention and control groups to assess compliance with drug medication and the patients' quality of life. The prohibition of a neutral dietary constituent as control advice might have been acceptable in the short term (for example, ice cream could have been prohibited). However, such advice would not have carried much conviction and the day-to-day dietary habits of the patients would have been largely unaltered. To give the control group major and

therefore comparable advice (such as a low-cholesterol diet) would not have provided an untreated control group. There may be no satisfactory method of performing a double-blind trial of dietary advice but the effect of a low salt intake can be determined by advising all patients to take the diet and add either salt tablets or placebo tablets according to whether they were in the control or intervention group.

8.8.2 The protocol differs between the intervention and control groups in order to save time or money

A theoretical example may be employed to illustrate a faulty protocol that requires the intervention and control groups to be treated separately. A new drug has to be tested against placebo treatment and blood tests and blood pressure measurements are required for the active treatment group but only blood pressure measurements in the placebo group, as a considerable amount of data has shown that placebo treatment does not affect the blood tests. However, taking blood in the actively treated group may raise the blood pressure in that group either because the blood test precedes taking the blood pressure or because the patient anticipates the blood test. No such rise will be observed in the group given placebo and whenever possible the protocol for the actively treated and control groups must be identical.

8.9 THE USE OF THE DOUBLE-BLINDING TECHNIQUE

The double-blinding or double-masking technique is very important and should be used, whenever possible, in all randomised controlled trials. Knowelden [99] has commented,

When everyone is in the dark, subjective measures can be used with confidence as there can be no bias introduced by patient or observer. If an observer had to diagnose a slight paroxysmal cough in a child known to have been vaccinated against whooping cough he might, because of bias in favour of the vaccine, decide it could not be pertussis. What is probably more likely than such cheating is that the clinician would attempt to compensate for his known bias and label the case one of whooping cough against his better judgement.

Double-blindness, however desirable, is not always possible and many trials have been performed without any blinding or with only masking of the patients (single-blinding). We must discuss how double-blinding can be achieved, when it cannot be utilised, why it is often not attempted, and what may happen if it fails.

8.9.1 How to achieve double-blinding

The following steps will lead to double-blinding:

1. The patient must understand that the trial is double-blind and give written informed consent.

MP-100
14148

2. The control treatment and experimental treatment must be identical. If tablets are used, they must be the same size, shape, colour, taste, and smell.
3. The exact nature of the treatment being given must be held in a secure position and be immediately accessible in an emergency. The treatment is often written on a card and enclosed in an opaque, sealed envelope. When the treatment is a drug these envelopes may be conveniently held in the pharmacy responsible for dispensing the treatment.

8.9.2 When masking is not desirable

Blindness is not desirable when it would result in unacceptable suffering or discomfort for the patient. The taking of dummy tablets will not be expected to produce any discomfort but sham operative treatment or dummy treatment involving repeated injections would do so. In the early Medical Research Council trial of antituberculosis therapy [3], it would not have been desirable to inflict dummy injections on the control group for several months. Similarly, although placebo operations have been performed (section 3.8) it is only possible to perform sham operations in humans in very exceptional circumstances. These reservations do not apply to animal experiments where sham operations are the rule.

8.9.3 When masking is difficult

Double-blinding techniques may be difficult when the treatment has obvious effects or when the treatment involves a change in life-style.

8.9.3.1 The treatment has an obvious effect

Certain drugs have specific actions which, if detected, would enable the patient or observer to break the treatment code. Examples are given by the beta-adrenoceptor blocking drugs that lower the pulse rate, and diuretics that usually result in an increase in serum uric acid. An observer could use this information to break the code and steps should be taken to ensure that he does not become aware of these results (section 8.5).

8.9.3.2 The treatment involves a change in life-style

Dietary change is an example of an alteration in life-style where masking is difficult. Unlike trials of drug treatment the advantages of double-blind trials of dietary advice are balanced by a number of disadvantages. In a double-blind trial, items of diet have to be provided by the investigator, some of which contain the dietary constituent under investigation and are given to one intervention group and some that appear identical but do not contain the constituent and are given to a second group. An example is given by the National Diet Heart Study [100] in which saturated fats were the dietary constituent under investigation.

Table 8-2 lists the advantages of double-blind and nonblind trials of dietary advice, assuming that the trial end points are not affected by masking, being

Table 8-2. Advantages of double-blind and nonblind trials of dietary advice. With a double-blind trial the subjects are given certain items of food whose composition is unknown.

Double-blind dietary study	Nonblind dietary study
1. Allows effect of diet to be assessed over and above nonspecific effect of dietary advice.	1. Assesses combined effect of dietary advice and taking the diet.
2. Changes in diet in control group less likely to occur (less contamination of control group).	2. Control group has an unrestricted diet.
3. Easier to secure unbiased ascertainment of all end points.	3. Easier to administer the trial and less expensive.
4. Easier to secure motivation of participants in control group (no difference in dropout rates).	4. Fewer ethical problems (patients and physicians know the treatment).
	5. Monitoring of blood or urine may be used to improve compliance in intervention group.

either easily ascertained (for example, death) or determined in a blind manner (for example, measurements of serum cholesterol). A nonblind trial of dietary advice tests the combined effect of diet and increased medical attention. Those who receive dietary advice must have greater contact with the research workers and this may affect the subjects' well-being in a nonspecific manner. The subject is aware that he has altered his diet and an increased attention to his food may either improve or reduce his sense of well-being. In a double-blind trial the intervention and control group will be equally affected by nonspecific dietary changes whereas in a nonblind trial the control group may have an unrestricted diet. However, in a nonblind trial some of the control group may alter their diet in a similar manner to that of the intervention group. If the control group hears about the intervention and alters its diet, the group is said to be contaminated.

As in randomised trials of other treatments double-blindness should ensure that the trial end points are determined similarly in the different groups and that the dropout rates are not affected by knowledge of the treatment. These advantages are balanced by a great increase in complexity and expense, possible ethical problems, and the fact that adherence to the diet cannot be closely monitored by blood or urine tests as these results, although useful in identifying noncompliant patients, will allow the treatment to be determined by the investigator if employed to improve compliance.

8.9.4 Strategies to be employed when double-blindness is not possible

When a treatment cannot be provided blind or when it produces obvious effects it may not be possible for one investigator to conduct a masked trial on his own. If the effects of treatment are determined from laboratory investigations the trial can be designed so that laboratory results are withheld from the

investigator. When a clinical measurement would detect the treatment, one investigator can determine the result and a second can be responsible for the care of the patients. Measurements can also be limited to those that are not open to observer bias (hard end points).

8.9.4.1 Blind the investigator to certain laboratory results

An example of this procedure may be provided by considering a trial of the blood-pressure-lowering effect of tienilic acid, a drug that markedly reduces the serum uric acid. During a double-blind trial the investigator measured blood pressure and interviewed the patients but the results of serum uric acid measurements were withheld from him [101]. It was planned that if a serum-uric-acid result was grossly abnormal and action required, the investigator should be told and the treatment code broken.

8.9.4.2 Two investigators: one to assess the patient and the other to provide treatment

Two investigators may have to be involved when therapy cannot be provided blind or when a treatment affects a measurement in a constant manner. One investigator provides the treatment or takes the measurement and the second assesses the end points of the trial. An example is provided by a trial of dietary advice where one investigator gives the therapeutic advice and a second assesses the outcome in terms of weight or blood pressure without knowing the treatment allocation. Similarly, in a trial of medical versus operative treatment of peptic ulceration (for example, by cutting the vagus nerve), it may be possible for a second observer to assess the results of the treatment from radiological plates or photographs of the ulcer. Care would have to be taken that such materials did not reveal whether or not surgery had been performed. It is possible that a moving image of barium past the ulcer may enable a radiologist to determine whether an operation had been performed or not.

In certain circumstances, the person providing the treatment may be masked and the person taking the measurements may not. Such an arrangement would be appropriate in a trial of the antihypertensive effect of increasing doses of a beta-adrenoceptor blocking drug compared with a different antihypertensive drug. Beta-blocking drugs lower both pulse rate and blood pressure and if the investigator knows the pulse rate, blindness would be lost. One investigator should measure blood pressure and pulse rate. The second investigator should remain blind and be given measurements of blood pressure but not pulse rate. He would then be responsible for prescribing increased doses of the drugs and assessing any side effects of treatment.

8.9.4.3 Hard end points

When masking is not possible, the analysis may be confined to so-called hard measurements. A hard end point is one that is not open to biased ascertainment; the best example is the fact of death. When survival is to be determined blindness is not necessary. However, although the fact of death may be incon-

trovertible the cause of death may be open to doubt. For example, sudden death may or may not be due to myocardial infarction and it may be difficult to distinguish cardiac from respiratory death. The difficulty is compounded when the investigator only knows the cause of death as written on the death certificate. These certificated causes are often not supported by postmortem examinations and are subject to error. Moreover, if an observer believes that a particular treatment prevents, or causes, a specific cause of death, this preconception may lead to a biased assessment of the cause of death.

8.9.5 The disadvantages of blinding

It is possible that masking will adversely affect patient recruitment and investigator participation. Identical control treatment will have to be provided at extra expense and it may be difficult to ascertain the exact treatment in an emergency. With drug treatment, it will be necessary for the patients to take extra tablets or capsules and labelling errors may occur. Lastly, masking may require additional personnel.

8.9.5.1 Fall in patient recruitment

When patients are informed that they and the investigator are unaware of the treatment they are to receive, some may be dissuaded from taking part in the trial. Similarly, they may dislike the idea of taking placebo tablets. In my experience this has only rarely been a problem.

8.9.5.2 Failure of investigators to take part in the trial

Potential investigators in multicentre trials may be unwilling to take part if a protocol is double-blind, although there is no data to substantiate this claim.

8.9.5.3 Provision of identical control treatment

Making tablets that are identical in appearance, touch, and taste may be difficult. The problem may be simplified by making identical capsules to contain the active drug or the placebo. However, if an active pharmaceutical compound is dispensed in a capsule, this formulation will differ from that of any tablets normally available and may affect the bioavailability of the drug.

If a control tablet is manufactured of an identical size, colour, and shape it may still smell and taste differently or differ in some other way from the experimental treatment. There is a story of a patient who reported that he knew his tablets had been changed at the last visit to the clinic as the new tablets were difficult to flush down the toilet! It is to be hoped that most tablets will not be disposed of in this or any similar manner.

It may prove possible to make an acceptable placebo for an active treatment but not to make two active treatments identical. In this instance, the patients can be asked to take two sets of tablets throughout the trial: one representing treatment A (active or placebo) and one B (active or placebo). This procedure has been called the double-dummy technique and may be useful in a cross-over

trial. In the first phase of the trial a patient takes one active drug and the placebo corresponding to the second drug, and in the second phase the active second drug is taken together with a placebo copy of the first drug.

8.9.5.4 *The patients have to take more tablets or capsules*

In a drug trial, taking a placebo always increases the number of tablets or capsules to be consumed. This is acceptable in order to make an unbiased comparison of an active treatment with an inactive treatment, rather than a test of an active treatment against no treatment at all. However, extra tablets may be required to preserve blindness when a placebo is not required, as with the double-dummy technique discussed previously.

8.9.5.5 *Expense*

When identical tablets or capsules are provided, it is expensive to manufacture, label, and distribute them. Dispensing may also be more complex in a double-blind drug trial.

8.9.5.6 *Errors in labelling*

In the labelling and distribution of identical tablets, great care must be exercised. The containers are only identified by code numbers or letters and the labels have to be checked and all stages carefully documented. I know of two trials where errors of labelling have occurred so that active treatment has been given instead of placebo and vice versa. Fortunately the errors were limited to only one or two patients. In one trial the errors were discovered when patients were checked for compliance: a patient on placebo had active drug metabolites in his serum when he had had no access to any active drug. In the second trial a patient on large doses of placebo suddenly received large doses of an active antihypertensive drug. The resulting side effects revealed the error.

8.9.5.7 *Difficulties in breaking the treatment code when necessary*

As discussed earlier, the code must be available for consultation in an emergency; it should be held in sealed envelopes, one envelope for each patient, and accessible 24 hours a day. The codes can only be held centrally if a coordinating office is manned day and night and otherwise must be held at a convenient place (for example, in the hospital pharmacy). The subjects in the trial should be given details of whom to contact in an emergency and these persons must know the location of the codes.

8.9.5.8 *The provision of extra personnel*

The need for two investigators to ensure double-blindness (see section 8.9.4) may be both difficult and expensive.

8.9.6 **The consequences of a breakdown in double-blindness**

The breakdown of double-blindness may affect the observations made in a trial and have serious consequences. In one trial of vitamin C against placebo

the active compound reduced the frequency, duration, and severity of the common cold [102]. However, this result was contrary to the results in other trials and the subjects were asked which tablets they thought they were taking. The subjects guessed correctly more frequently than they should have done by chance alone. This was a worrying finding, as those who knew they were on placebo may have reported more colds and those who realised they were on active treatment, less colds. Alternatively, a positive answer could have been masked if those on vitamin C, as a consequence of this knowledge, had not eaten as much fresh fruit and vegetables as those who thought they were on placebo. As the trial was to determine whether or not the vitamin had a nonplacebo effect on the incidence of colds, the wrong answer may have been obtained.

It is good practice to ask each patient about the treatment they guess they are taking. In the National Diet-Heart Study [100], participants were asked to purchase special dietary foods with different fat contents and the trial was conducted in a double-blind manner. The participants were asked about the amount of fats in their diets and 43 percent in each dietary group considered that their diet included a large reduction in total fat. There had been no loss of double-blindness and dropout rates were independent of diet as were the proportions volunteering for further study. Similarly, the doctors and nutritionists involved in the trial were asked to specify the diet that had been assigned to the individual patients. The correspondence between the actual diets and their guesses was no better than would be expected by chance.

A consequence of the breakdown of blindness has been examined in a trial where the investigators were apparently able to identify the active treatment. Heaton-Ward [103] performed a double-blind trial to assess the effect of a monoamine oxidase inhibitor (Niamid) on the activity and behavior of mongol patients. The observers were told that the trial was of a cross-over design but at the time of cross-over the same treatment was continued. The observers reported an initial improvement in the actively treated group but not in the placebo-treated group. However, after the supposed cross-over, they reported a deterioration in those who had initially improved and an improvement in those they first imagined had not improved. The objectivity of the observations appeared to be in some doubt. Abraham [104] termed this failure of blinding and its result the *Heaton-Ward effect*, after the author of this trial. The trial was concerned with activity and behaviour and in this field subjective impressions bedevil the interpretation of results (see section 16.6). A Medical Research Council Trial into the effects of antidepressant drugs [105] showed that a breakdown of blindness due to side effects led to differences between treatments being observed in subjective assessments. The effects of treatment were not confirmed by a more objective reduction in the length of stay in hospital.

Double-blinding is very important, and Beecher commented, “. . . there is evidence in surgery as in other fields, that the enthusiast actually gets results which are better than those of the sceptic” [38]. In a double-blind trial both the

subjects and investigators should be interviewed to determine whether or not masking has been preserved.

8.10 THE USE OF PLACEBOS

When effective treatment is available this is usually employed in the control group. However, in the short term, or when no effective treatment has been discovered, placebos should be employed as control treatment whenever possible (section 3.8).

8.10.1 Why should placebos be used?

Bradford Hill stated, when discussing the clinical trial, "To some patients a specific drug is given, to others it is not. The progress and prognosis of these patients are then compared. But in making this comparison in relation to the treatment the fundamental assumption is made—and must be made—that the two groups are equivalent in all respects relevant to their progress, except for the difference in treatment" [1]. As discussed in sections 8.4 and 8.5, the use of placebo treatment in the control group ensures that any difference between the actively treated and the control group is due to the active constituent employed in the trial and is not a nonspecific effect of giving any treatment.

8.10.2 When should placebos be used?

Placebos should *not* be used as control treatment when there is definite evidence that withholding available treatment may be detrimental to the patient's health. Other reasons have been advanced for not employing placebo treatment: the increase in cost of the trial due to placebo materials, distribution, coding, documentation, and the increase in the work load of the investigator (section 8.9.5). However, when there is no treatment of proven worth, a placebo-controlled trial is to be preferred to one using an untreated control group. In some trials the use of a placebo is obligatory; in other trials it is advisable but optional.

8.10.2.1 Essential uses of placebo

Placebo treatment is essential in trials of anxiolytic, hypnotic, and antiinflammatory drugs. The use of a placebo regime allows the day-to-day variation of subjective sensations such as pain to be measured together with any spontaneous improvement with time or as a nonspecific response to tablets.

8.10.2.2 Important uses of placebo

Placebos are often employed in chronic conditions such as hypertension and diabetes mellitus when available treatment is known to correct the pathophysiological abnormality but has not been shown to reduce mortality or morbidity. Three examples where active treatment has not been proven to be of benefit to the patient's health are mild hypertension (diastolic pressure 90–99 mm Hg), benign hypertension in the elderly, and mature-onset diabetes. The long-term treatment of these conditions with a placebo is justified as no

known treatment confers definite benefit in terms of reduction in mortal or morbid events. In hypertension, the use of a placebo is particularly important as the control group may exhibit a reduction in blood pressure owing to the reassurance offered by supposedly active treatment. However, a specific placebo effect in reducing blood pressure has not been proven and the reduction in pressure may reflect the process of familiarisation with medical attendants and clinical surroundings.

8.10.2.3 Other uses of placebos

In the chronic diseases discussed previously, placebo treatment is often employed in the short term to determine a baseline level of blood pressure or blood sugar even when the patient is known to require active treatment in the long term. A frequent example of the use of a placebo is in young patients with moderate or severe hypertension. The investigators know that active treatment must be given eventually, but the treatment under trial is to be compared with a baseline untreated period. Taking a placebo ensures that the control period is identical with the period of active treatment. Two strategies partially resolve the ethical problems: the length of the placebo treatment is limited to a short period of 3–8 weeks, and the placebo treatment is stopped if blood measurements exceed an arbitrarily defined level (see section 3.8).

8.10.3 How placebos should be employed

8.10.3.1 The agreement of the subject must be obtained

The patient must agree to take either the placebo or the experimental treatment. Fortunately, many patients are prepared to serve as experimental subjects and contribute to the common good. Patients presenting to a doctor are usually willing to adopt the advice they are given and if the doctor suggests taking part in a placebo-controlled clinical trial, the patients tend to accede to the request. They may be unwilling to make an independent choice between entering or not entering the trial. Such indecision might be analogous to the situation where an airline pilot, with a faulty aeroplane, asks his passengers whether he should go on or turn back. The doctor's advice will be taken and the trust of the subjects must not be abused.

All relevant information about the trial must be provided for the subject, including the following:

1. The fact that a placebo is being used in the trial and the patient may receive it.
2. Whether or not the trial tests a concept that could lead to a benefit for the patient.
3. That any new treatment has been adequately tested in the laboratory and may be an advance over existing therapy.
4. That the probable risks involved from taking the new drug or placebo are

The patient should be given a written copy of this information to facilitate full comprehension and be asked to provide written consent to taking part in the trial. Such consent does not reduce the investigator's responsibility but does provide proof that the patient read about the trial and agreed to take part on the basis of the information. It also proves that the investigator discussed the trial with the patient (section 3.7).

Bradford Hill questioned whether the patients should be told that they may receive a placebo and he wrote, "Having made up your mind that you are not in any way subjecting either patient to a recognized and unjustifiable danger, pain or discomfort, can anything be gained ethically by endeavouring to explain to them your own state of ignorance and to describe the attempts you are making to remove it? . . . Once you have decided that either treatment *for all you know* may be equally well exhibited to the patient's benefit, and without detriment, is there any real basis for seeking consent or refusal?" [30]. Many would support this view for a trial of acute treatment in a stressful situation (for example, on admission to a coronary care unit with acute myocardial infarction). Under less stressful conditions and with long-term treatment informed consent should be obtained.

8.10.3.2 Single or double blind?

The trial employing a placebo must be single or double-blind (section 8.9), and the placebo treatment must be identical to the active treatment. If a tablet, it should look, smell, feel, and taste the same.

8.10.3.3 Randomise the order of treatment in a cross-over trial

A placebo run-in period is often employed in randomised trials to ensure that only patients with certain characteristics enter the trial. An initial period on placebo treatment allows patients to be excluded if they do not have, for example, a persistent elevation in blood pressure, blood sugar, or serum cholesterol. Subjects who do not comply with the trial protocol may also be identified during this period. However, when a period of placebo treatment is to be compared with an interval on active treatment in the same patient, the order of treatment must be varied, using random allocation. Often the placebo treatment is given first and this error may not be immediately obvious when the investigators are comparing two treatments with placebo. It is not acceptable to report a placebo-controlled randomised double-blind cross-over trial of two treatments when the placebo period always precedes the randomised part of the trial. In a trial of two antihypertensive drugs the reader would assume that the baseline blood pressure was accurately determined by a randomised period of placebo treatment during the trial. As the average blood pressure tends to fall throughout treatment, the effect of the drugs in lowering blood pressure may be exaggerated. Baseline measurements should be obtained double-blind during the trial with the two active treatments and the placebo treatment given first, second, and third with equal frequency.

8.10.4 Disadvantages of the use of placebos

The disadvantages arising from the use of placebos overlap those given for double-masking (section 8.9.5). In addition, the use of a placebo may disrupt the doctor-patient relationship and produce medico-legal problems.

8.10.4.1 Disruption of the doctor-patient relationship

Bradford Hill stated, "The doctor will also wish to consider the doctor/patient relationship. Harm may be done if the public comes to believe that doctors are constantly using them as guinea-pigs. In exhibiting new treatments they are, it is my belief, doing that willy-nilly, but the public does not realize it. But they need not go out of their way to make it obvious by an unnecessary use of dummy pills" [30]. But what is an unnecessary use? The provision of an exact control group by the use of placebos is often essential.

8.10.4.2 Medico-legal problems

It is possible to argue that if a drug has been shown to be beneficial in one clinical trial against placebo, an investigator repeating the trial is knowingly placing at risk any patient treated with placebo. If patients in the placebo group suffer harm, the investigator may be sued for damages on the grounds of negligence. Dollery [86] stated, "Lawyers appear to have little time for the contention that there is uncertainty about the efficacy when only one or perhaps two trials have completed. They do not understand the concept of statistical uncertainty and are accustomed to resolving doubts about factual uncertainty in the courts."

An investigator may also run into problems when he employs an experimental treatment and there is no definite evidence that it will be successful. If the experimental treatment subsequently proves of benefit in later trials, the courts may examine the original trial and find in favour of subjects in the placebo group who have suffered harm. Dollery provides the example of trials of active antihypertensive treatment in patients with an untreated diastolic blood pressure of 90–105 mm Hg. This group has only mild hypertension and the original trial, the Veterans Administration Co-operative Study on Antihypertensive agents [37], found a 35 percent reduction in complications in this group with active treatment; but this did not achieve statistical significance. Moreover, the trial included a high proportion of patients who had already suffered a complication of hypertension and patients were only included if the high blood pressure was maintained during a hospital admission. For these and other reasons further trials of antihypertensive therapy have been started for mild diastolic hypertension in young and middle-aged patients [106]. One trial has been completed and shown a benefit in patients with a diastolic pressure over 100 mm Hg but not 90–99 mm Hg [107]. If all trials subsequently find a benefit from active treatment when the initial diastolic pressures are between 100 and 104 mm Hg, can such patients take legal action when they have sustained a disabling stroke while on placebo in the original or later trials?

If the courts find in favour of the patient, will they realise that they do this with the assistance of hindsight? When the first trial was started, for all the investigators knew the patients treated with placebo may have been the fortunate ones as the result of such trials has often been in favour of the placebo group. However, many would claim that patients who suffer from taking part in randomised trials should be compensated (section 3.2). In this event the compensation payable should not be excessive as pharmaceutical companies and other organisations may be unwilling to support placebo-controlled trials. Imagine the problem if trials of new treatments are conducted only against a supposedly active treatment, the active treatment being of no use. There could be a proliferation of useless drugs all being as good as the active (useless) treatment.

8.10.5 How many positive trials are required against placebo?

How many trials of active versus placebo treatment have to show a positive result before the results can be accepted and further use of placebo prohibited? Only two trials of antihypertensive treatment were performed before it was accepted that middle-aged men with a sustained diastolic pressure over 105 mm Hg should be treated and not exposed to long periods of placebo treatment. However, when trial results are contradictory, as with the use of anticoagulants following myocardial infarction, a large number of trials may have to be performed. The International Anticoagulant Review Group examined nine trials where proper control groups were employed and concluded that the benefits from anticoagulants were of the order of a 20 percent reduction in mortality in men and zero in women [52]. The medical profession has not considered this to be a worthwhile gain in men and the treatment has not been employed in most countries (see section 19.4). On the other hand, a more recent trial showed a significant reduction in myocardial reinfarction following prophylactic treatment with sulphapyrazole [48]. A positive result may occur by chance in up to 5 percent of trials (according to the level of significance achieved in the trial) and it would appear reasonable to require two positive trials before accepting any results as conclusive. A second trial, including placebo treatment, may therefore be started in order to confirm the initial finding.

8.11 CONCLUSIONS

The results of a trial may be expected to be free of bias when randomisation has produced an equivalent control group. The control group must provide an estimate of any placebo effect, an estimate of any time trend in the condition being investigated and of the effect of regression to the mean.

Noncompliance with therapeutic advice may alter the response to treatment but patient and observer bias may be removed by single- and double-blinding respectively.

Great care must be taken in the analysis of the results of the trial in order to

avoid introducing bias, especially when considering subjects who withdraw from the trial.

The method of conducting the trial is very important and double-masking and placebo control should be introduced where possible. Both these techniques can present difficulties in the form of complexity, expense, and even litigation. If double-blinding, placebo control, and randomisation are not employed the consequences are likely to be more serious than any adverse results from using them.

9. THE VARIABILITY OF RESULTS

Even when the results of a trial are free of bias, the measurements on which they are based are still subject to random variability. Often the variability can be both estimated from its repeatability and reduced by quality control. The measurements may vary between subjects and this variability can be reduced by employing the within subject cross-over trial design. It must also be remembered that a measurement may be very repeatable but either not provide the correct result (low accuracy) or measure something other than that intended (low validity).

9.1 REPEATABILITY

Repeatability is the level of agreement between replicate measurements. The level of agreement must be determined for measurements made in the same subject in order to exclude differences between persons. The variability may result from observer error, machine error and, when repeatability is estimated at widely spaced intervals, from alterations in the true measurement with the time of estimation.

9.1.1 Repeatability of a continuously distributed variable

The repeatability of a continuously distributed variable is estimated by the standard deviation of repeated measurements. The difference between the av-

erage result of these measurements and the true result should also be calculated when the latter is known. The standard deviation of repeated measurements has been termed the *precision of the measurement* [108], although when the deviation is high it will be less precise. The difference of a result from an expected value has been termed the *accuracy of the measurement*. The standard deviation is often divided by the mean and expressed as a percentage known as the coefficient of variation.

Obviously, it is important to make very repeatable measurements during the course of a trial and great difficulties have arisen when this has not been the case. The World Health Organisation and the Center for Disease Control, Atlanta, have been involved in assessing the repeatability of serum cholesterol measurements around the world [108]. Cooper has reported these results and found that, although the standard deviation of the various methods was satisfactory, some methods gave high readings of cholesterol and poor accuracy. The two methods causing the most problems were the direct method of cholesterol estimation and the chloride method. It is worth emphasising that if a laboratory changes its method of analysing cholesterol or any other substance during the course of a trial, the measurements may be altered with serious consequences. The new method may lack precision or differ in accuracy when compared with the old method.

Figure 9-1 illustrates how a biochemical measurement on the blood can be checked three times during a clinical trial. This assessment will be most important in a long-term trial and, when a placebo group is employed, repeated measurements in this group will not be affected by treatment. In the example each subject has blood taken on three occasions and the blood is split into three samples. Biochemical estimations on the samples give the results designated X . The results for the first subject have a suffix 1 (X_1); his results for the first occasion are designated X_{11} and results for the first sample on that occasion X_{111} . The subjects are numbered 1, 2, 3, n in the first suffix which we shall call i ; occasions 1, 2, and 3 are given the second suffix; and samples 1, 2, and 3 the third suffix. The average for a suffix in the figure is given by a point. The average result for all measurements is given by $\bar{X}_{...}$, the average for subject 1, $\bar{X}_{1..}$, and for occasion 1, $\bar{X}_{.1.}$.

9.1.1.1 The repeatability on one occasion

The within-specimen variance for subject 1, occasion 1, is given by

$$\sum_{s=1,3} \frac{(X_{11s} - \bar{X}_{11.})^2}{r-1}$$

where $\sum_{s=1,3}$ is an instruction summing the squared differences for r samples numbered in this example 1 to 3 (s = sample).

OCCASION SPECIMEN TAKEN										
	1			2			3			Average
	Split Sample			Split Sample			Split Sample			
	1	2	3	1	2	3	1	2	3	
Subject 1	x_{111}	x_{112}	x_{113}	x_{121}	x_{122}	x_{123}	x_{131}	x_{132}	x_{133}	$\bar{x}_{1..}$
Subject 2	x_{211}	x_{212}	x_{213}	x_{221}	-	-	-	-	-	$\bar{x}_{2..}$
Subject 3	x_{311}	x_{312}	x_{313}	-	-	-	-	-	-	$\bar{x}_{3..}$

Subject n	x_{n11}	x_{n12}	-	-	-	-	-	-	-	$\bar{x}_{n..}$
Overall mean	$\bar{x}_{.1.}$			$\bar{x}_{.2.}$			$\bar{x}_{.3.}$			$\bar{x}_{...}$

Figure 9-1. The assessment of repeatability during a clinical trial (see text). The data provide information on the precision and accuracy of measurements on each occasion.

For all subjects and occasion 1 the precision of the measurement

$$= \sqrt{\left(\sum_{i=1, n} \sum_{s=1, 3} \frac{(X_{is} - \bar{X}_{i.})^2}{r-1} \right) / n}$$

where the squared differences are summed over n subjects and three samples.

In the same manner the repeatability of the measurement should be estimated for occasions 2 and 3 and compared with the initial measurement of repeatability. A worked example is given in Appendix 9.5 to compare the blood pressure measurements on two prototype samples of a new kind of sphygmomanometer with one mercury sphygmomanometer. The prototypes did not prove to be acceptable.

The accuracy of the measurement can be gauged by comparing the mean for occasion 1 ($\bar{X}_{.1.}$) with the means for the other two occasions or, similarly, the

		FIRST OBSERVER	
		Diagnosis +	Diagnosis -
SECOND OBSERVER	Diagnosis +	a	b
	Diagnosis -	c	d

Figure 9-2. How the data may be presented to demonstrate the agreement between two radiologists examining the same films; a, b, c, and d are the number of pairs where they agree or disagree.

average result for a standard machine with the average for the machine under investigation.

Another method can be employed in a long-term trial to examine for a drift in the results. For each measurement during the trial the difference is calculated between this figure and the initial average results. The positive differences should balance the negative ones, giving a cumulative sum of differences near to zero. The cumulative sum of differences is calculated during the course of the trial and plotted against time. A deviation from zero increasing with time will indicate a change in accuracy. The graph has been termed a *CUSUM plot* and has been employed in monitoring laboratory performance.

9.1.2 Repeatability of qualitative data

Qualitative data are not continuously distributed and include diagnoses, symptoms, and clinical states given a discrete value (for example, zero when the condition is absent and one when present). Repeatability of qualitative data is most easily measured by comparing two estimates (for example, when radiologists report a diagnosis). To estimate between-observer repeatability two radiologists may be asked to examine the same radiographs and report their opinions. If one observer is given the same films to examine on two occasions then within-observer repeatability can be estimated.

Figure 9-2 illustrates how the data may be presented, the columns giving the opinions of the first observer and the rows the results from the second observer. The letter "a" represents the number of positive responses in both

FIRST ASSESSMENT

a) 20% prevalence

		+	-	TOTAL
S E C O N D	+	10	10	20
	-	10	70	80
	TOTAL	20	80	100

$R = \frac{80}{100} = 80\%$

b) 2% prevalence

		+	-	TOTAL
S E C O N D	+	1	1	2
	-	1	97	98
	TOTAL	2	98	100

$R = \frac{98}{100} = 98\%$

Figure 9-3. The measure of agreement, $R = (a + d)/(a + b + c + d)$, is not independent of prevalence being 80% in the first example and 98% in the second.

assessments; "d" represents the number of negative responses on both occasions; "b" assessments were negative with the first observer but positive with the second observer; and "c" assessments were positive with the first observer and negative with the second. The agreement has been measured by the expression:

$$R = \frac{a + d}{a + b + c + d}$$

However, this index varies with the prevalence of positive findings. Figure 9-3 shows the result of calculating R in a condition with a 20 percent prevalence in both assessments and an equal division between the positive and negative cases, and similarly the result when the prevalence was only two percent. The

index (R) was 80 percent with a 20 percent prevalence and 98 percent with a two percent prevalence.

Two indices of repeatability have been published that are independent of prevalence:

$$\text{Index 1} = 1/4 \left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{c+d} + \frac{d}{b+d} \right)$$

This index can be calculated as long as the sum of a row or a column is not zero [109].

$$\text{Index 2} = 1 - (bc \div ad).$$

This index is also independent of prevalence but cannot be calculated if any cell is zero [110].

Repeatability indices can be used to compare diagnoses made by different observers, the same observers on different occasions, or by different means (for example, electrocardiographic evidence of myocardial infarction compared with enzymatic data). The indices have been used to compare the responses of patients to self-administered questionnaires on two different occasions. Table 9-1 provides some examples of the repeatability of symptoms in 55 normal subjects over a ten-month period as measured by index 2. Index 2 should be zero if there was no agreement between the responses to the self-administered questionnaire and 1 if there was full agreement. The repeatability measurements were all .95 or greater for single questions but were lower when more than one question led to a conclusion. For example, the subjects were asked if they had headaches and whether these headaches occurred on waking. The repeatability index for the pair of questions eliciting the response, waking

Table 9-1. Repeatability of questions on symptoms as determined by two self-administered questionnaires, ten months apart.

Symptom	Repeatability Index
Single question	
Faintness	0.95
Headaches	0.98
Blurred vision	0.99
Depression	0.95
Nocturia	1.00
Double questions	
Faintness on standing	0.94
Waking headaches	0.87

Index = $1 - (bc \div ad)$. (See Text.)

headache was only .87. However, it appeared that the single questions had an acceptable repeatability. Eliciting data on symptoms is discussed further in chapter 16.

9.1.3 The relationship between bias and repeatability

If a measurement is very variable, with a low repeatability, replicating measurements and taking the average may overcome the problem. In drug trials on hypertensive patients, repeated blood pressure measurements may be taken before and after treatment. By including sufficient observations and enough patients in the trial (see chapter 10) the results may be acceptable as a large number of patients and readings will give an accurate mean with low standard error. With smaller numbers poor repeatability due to random error may obscure the beneficial effects of drugs or relationships between measured variables, but *bias* may produce a result that is incorrect. In within-patient cross-overs (section 9.3) the order effect of giving the drug represents a possible bias. An example was provided in section 8.10.3 where a subject may become more used to a blood pressure measurement between the first and second stages of the trial and the blood pressure may fall, not due to therapy but due to an order effect. If two drugs are being compared one can be started first in a random half of patients and the other first in the remaining subjects. If the order effect is not balanced out by this strategy the results will be biased in favour of the second treatment. Poor repeatability may lead to an answer not being obtained but bias may lead to the wrong answer.

9.1.4 The effect of regression to the mean on repeatability

Regression to the mean was defined by Francis Galton in 1886 as "each peculiarity in a man is shared by his relatives, but on the average to a lesser degree." His definition meant that a very tall man was likely to have shorter brothers and shorter sons and a very short man taller brothers and taller sons [111]. The concept has been extended to describe the change in variables through time that results from selection at the start. Many examples of regression to the mean exist and a good example is provided when blood pressures are measured. Subjects with pressures initially higher than the mean tend to have lower pressures on the second occasion and those with starting lower blood pressures than the mean tend to be higher on the second occasion. Therefore, if you select a group of hypertensive patients on the basis of their high blood pressures, their measurements will certainly be lower on the second occasion owing simply to this regression-to-the-mean phenomenon. This fall in pressure and loss of repeatability should occur both in an actively treated and a control group and the provision of the control group will prevent the effect being ascribed to active treatment. Regression to the mean can also be limited to some extent by only including patients with a repeatedly high pressure, such patients presumably having a smaller tendency to drop to lower pressures.

9.1.5 The effect of repeated assessments on the prevalence of a condition

Higgins' Law states, "The prevalence of any condition is inversely proportional to the number of experts whose agreement is required to establish its presence." Let us assume that all observers have to make a certain diagnosis for it to be present. If one observer considers that 100 out of 1,000 subjects have angina, a second observer may also agree that 100 have angina but the latter 100 may only include 75 of the initial 100. Twenty-five of both the initial and second diagnoses were not in agreement, reducing the confirmed prevalence of angina to 75 in 1,000. The use of a third observer may eventually give a prevalence as low as 60 angina cases per 1,000. When conducting a long-term clinical trial, it is often essential to have more than one observer but one must realise how the prevalence of a condition, which may be an end point of a trial, may be reduced by repeated examinations.

9.1.6 Good repeatability does not imply validity

The validity of a measurement depends on whether or not it estimates what it is supposed to measure. For example, a measurement of blood sugar may be examined, a blood sample split in two, and the coefficient of variation for a single sample computed with very satisfactory results. However, if random blood sugars are being used to diagnose diabetes mellitus, within-sample repeatability may be satisfactory but between-occasion repeatability may reveal considerable variation. Moreover, a high random blood sugar may not be a good indicator of diabetes mellitus. It may vary according to the recent diet, being high after a carbohydrate meal and low after fasting. A single measurement may have a high repeatability but low validity when used as a diagnostic test in diabetes. A diagnostic test must have both a high validity and a high repeatability and rarely will a test have high validity and low repeatability.

9.2 QUALITY CONTROL DURING A TRIAL

In order for a trial to be conducted exactly as intended, the instructions in the protocol must be clear, the observers fully trained, and the progress of the trial monitored to detect deviations from the trial protocol and other errors.

9.2.1 A clear protocol

Writing a protocol is discussed in chapter 12. The protocol must be detailed and precise, specifying exactly how the trial is to be conducted so that the measurements may be made in a constant manner. For example, if blood pressure is to be measured, the protocol should give the details in table 9-2. For each heading on the left of the table, a statement must appear in the protocol and also in the final report.

Attention to detail is so important that a trial should usually be conducted for a pilot period when the protocol can be amended in the light of experience (section 14.1). Ferris and Ederer [112] have discussed protocol violations in

Table 9-2. The detail that may be required in a protocol for a trial where blood pressure is measured

Machine	Standard sphygmomanometer/random zero, etc.
Cuff size	State size of inflatable portion
Cuff-deflation rate	2 mm/second?
Circumstances	Measurement to be made in the doctor's office, laboratory, at home?
Position of patient	Lying, standing, sitting?
Relaxation of patient	Resting for 10, 5, 2 minutes?
Time of day	Morning, afternoon, evening? Certain number of hours after treatment?
Replication of measurements	If replicated, which reading(s) are recorded?
Measurement of diastolic pressure	Point of muffling or disappearance of sound?
Accuracy	Reading to the nearest 2 mm Hg?

ophthalmological studies, and table 9-3 reports some of these violations observed during a pilot trial where visual acuity was to be measured. Certain violations led to the clinic changing its practice, some to the protocol being changed and others to the development of special equipment to overcome the difficulties. Of particular interest was the fourth protocol violation, where the patients had to read a line of letters and were judged to be able to do so if they only got one letter incorrect. For the line 20/200, the charts included only one letter and in theory, all patients should have been judged capable of reading that line. New charts were constructed with at least four letters per line.

The violation 5.c was also of great interest. A standardised protocol for refracting patients and obtaining the visual acuity had been developed but the techniques were rarely followed. The greatest variability occurred when ophthalmologists were doing the examinations "A few . . . clearly did not know the study protocol at all." In many clinics technicians were trained to take over the measurements and they adhered strictly to the protocol. The authors suggested that the protocol way was the only way the technicians could make the measurements and therefore they used this method and got reproducible results. The training of observers is of great importance.

9.2.2 Training of observers

Observers must be trained to achieve both repeatable and accurate results. An observer who makes repeatable measurements can still make the measurements consistently too high or too low. An example of the necessity for training is given by the measurement of blood pressure.

9.2.2.1 Learning the technique

The observer has to be taught which machine to use, how to apply the cuff, as well as the requirements listed in table 9-2. Assuming that the observer is not deaf there are still difficulties in detecting the diastolic sounds. The fourth

Table 9-3. Protocol violations observed during a trial involving the measurement of visual acuity before and after refraction, and how the violations were dealt with [112]

Protocol requirement	Violation	Result
1. Visual-acuity lane 20 feet long.	Distance reduced: difficult for clinic to provide distance.	Protocol changed: 10-foot lane.
2. Front illumination.	One clinic provided rear illumination.	Clinic changed practice.
3. Illumination by tungsten spotlights.	a. These lamps not used (e.g., giving only 10% of specified lighting). b. Bulbs aged and illumination reduced. c. Uneven illumination.	Clinics changed practice. Specially developed visual-acuity boxes issued.
4. Visual acuity defined as that line on which patient gets not more than one letter wrong.	Charts issued had only one 20/200 letter. (Patients could not get this line wrong!)	New charts issued with four 20/200 letters.
5. Refraction.	a. Not done by trained personnel. b. Not done by blind observer. c. Measurement rushed and protocol ignored (or unknown).	Examiners had to be trained and receive certification of training. Technicians trained to take over this measurement from ophthalmologists. Replications done by site visitor.

Korotkoff sound is the point of muffling of sounds (when the sounds stop having a tapping character) and the fifth Korotkoff sound is when the sound disappears. Observers can be trained to recognise these sounds by listening to them while observing a film of a mercury column falling. The observers thus gain experience and are then tested until they make the simulated pressure measurements consistently close to a standard measurement. For real measurements of pressure a stethoscope can be adapted so that the trainee and the trainer can listen to the sounds at the same time and the training session can continue until consistent agreement has been reached.

9.2.2.2 Adherence to the protocol

The observer must be tested for adherence to the protocol. If the protocol requires blood pressure measurements to the nearest 2 mm Hg then the end digits 0, 2, 4, 6, and 8 should be recorded with equal frequency. When monitoring the quality of blood pressure measurement, a preference for an end zero is often observed and indicates failure to adhere to the protocol. However, digit preference is much less important than an overreading or underreading due to *observer bias*. Two machines, used correctly, should prevent this bias. A random zero sphygmomanometer is available where pressure is recorded and then the zero point determined and later subtracted [113]. Digit preference is

less obvious with this machine if zero measurements are made to the nearest 2 mm, but there is nothing to stop digit preference from the pressure's being read to the nearest 10 mm Hg. The London School of Hygiene and Tropical Medicine sphygmomanometer does not allow the observer to watch the mercury column while listening to the sounds, the descent of three columns being arrested by the observer when the changes in the sounds are detected [114]. The lengths of the mercury columns are then measured at leisure and observer bias should be negligible provided that the observer records the first reading and does not repeat the measurement (section 8.5). All measurements must be conducted according to a specific protocol and new instruments developed if necessary.

9.2.3 Should the same or different observers be employed?

Conventional wisdom suggests that if the same observer makes measurements throughout the trial and adheres to the protocol, the measurement technique will be constant and standardised. Alternatively two or more observers can be trained and tested to ensure that they make similar measurements. There are several disadvantages in having only one observer. The observer may be unable to make all the measurements (for example, through illness, holidays, or promotion to another position) and the substitute will be unlikely to have been as well trained as the original observer. In addition, it is difficult for another group to replicate a trial where much depends on the measurement technique of one individual. It would appear safest to train and employ at least two observers.

9.2.4 Checking of trial documents

All trial documents must be checked as soon as possible after completion. The design of forms to minimise errors in completion is discussed in section 13.8. At all steps of data collection errors must be avoided. If the data are to be entered in boxes then it must be clear where the decimal point is expected. Biochemical and other data must be entered in the units in which they are collected, as it is easier and less prone to error for a computer program to convert data from one set of units to another than for a busy clinician to do so in his outpatient department. The documents must be free of any ambiguity. For example, if the sitting phase IV (muffling of sound) diastolic blood pressure is required, this should be stated in the document as well as in the protocol. Finally, the forms should be designed with efficient data processing in mind but this constraint must not make the documents difficult to complete.

The documents for a controlled trial should not accumulate at a local centre until the end of the trial. Either the documents should be checked on the spot (see the functions of a monitor, section 9.2.5) or they should be sent directly to a coordinating centre. Whatever the method of monitoring, *quality control* must be employed to ensure that patients entering the trial are eligible and that the protocol is adhered to during the course of the trial.

9.2.4.1 Quality control of entry criteria

The initial records must be examined to confirm that the patients are truly eligible to enter the trial. It is surprising how many patients are entered into trials by experienced investigators and yet do not fulfill the entry criteria for the trial. If the patients are randomised after scrutiny of the initial records the proportion of ineligible patients who enter will be greatly reduced.

9.2.4.2 Quality control during the course of the trial

The documents should be designed so that a monitor can see that the protocol is adhered to; for example, if a tablet count of within ten percent of the expected number is necessary to determine compliance then the investigator should record the number of tablets rather than whether or not noncompliance was present. Patients often forget to bring their tablets with them but assure the investigator that they have taken their medication and he will be tempted to record full compliance without the objective evidence required in the protocol. Similarly a protocol may require three blood pressures to be taken and the last to be recorded. In order to check adherence to the protocol, the investigator should be asked to record all three measurements.

The documents should be checked both for consistency and errors. Certain information must be repeated on every form to ensure that they relate to the correct patient. Errors will include those of diagnosis, measurement, failure to withdraw patients when necessary, and also clerical mistakes.

9.2.4.3 Errors in diagnosis

It cannot be assumed that all diagnoses made during the trial are correct. Kahn and his colleagues [115] provided some very interesting examples of diagnostic difficulty in glaucoma and concluded that more training, more tightly written protocols, and an increase in the use of objective measurements are all required. It is also important that the criteria on which a diagnosis is based are recorded and not simply the diagnosis.

9.2.4.4 Error in measurement

Errors of measurement may be random or systematic. Both are important, and the latter (systematic error or bias) can produce a false result if it occurs in one treatment group more than another. Before commencing a trial, other than a small pilot trial, the investigator will have some knowledge of the variability of the measurements to be made. This information may be necessary for the calculation of the numbers required for a particular trial and will be required for quality control during the trial. The acceptable range of results for every measurement must be defined and the documents examined to detect results that lie outside these limits. The outlying measurements must be investigated further.

In quality control we are concerned with both the repeatability and validity

	Patients' answers	
	Yes (have been admitted with heart attack)	No (have not been admitted with heart attack)
Known to have been admitted with a heart attack.	a	b
Known to have been admitted with a different diagnosis.	c	d

Figure 9-4. The validity of self-administered questionnaire to determine whether or not a patient's recent hospital admission had been due to a heart attack. The letters a, b, c, and d are numbers of patients.

of the data. Figure 9-4 provides a theoretical example in which the validity of a question, "Have you been admitted to hospital with a heart attack during the last six months?" is examined. The question was asked of a group who had been admitted with myocardial infarction and a group that had been admitted with a variety of other medical conditions. Their answers are entered in the columns of the figure and cross-tabulated according to the known truth. If the total number of patients was n then the proportion of correct responses was $(a + d)/n$. Similarly the false negative rate can be calculated as $b/(a + b)$ and a false positive rate as $c/(c + d)$. Such a tabulation could be used to examine the validity of other data. For example, the vertical columns could be a normal or high random blood sugar and the horizontal rows the presence or absence of diabetes mellitus as determined by a glucose tolerance test. Similarly, electrocardiographic abnormalities, or serum enzyme changes, may be examined separately as tests for acute myocardial infarction provided a definitive diagnosis or gold standard can be provided from, say, coronary angiography.

9.2.4.5 Failure to withdraw patients according to the protocol

The criteria for withdrawing a patient must be stated in detail in the protocol and the documents completed on withdrawal must include the exact reasons to ensure that the protocol is adhered to. If a patient has to be withdrawn from a

trial when the serum creatinine exceeds a certain concentration, then the documents must request the results of serum creatinine determinations. When patients are withdrawn from a trial for reasons outside the protocol this occurrence must be detected during the quality control procedures and the investigator interviewed to ensure that the protocol has been understood.

9.2.4.6 Clerical errors

Clerical errors are mainly those of transferring information and four have been defined [116]: person to document; instrument to document; document to punch cards; and computer output to report. We can add another transfer error, typing mistakes in any of the draft or final reports.

The solution to these errors is repeated checking. Two persons should be responsible for the completion of all documents and punching of cards must be verified by repunching. In the latter process the card is rejected by a verifier when the original holes in the computer card do not coincide with the second keyboard input. Despite all care, errors still occur and all data have to receive two examinations: range and consistency checks. Range checks ensure that the results lie within the range of possible answers and consistency checks examine the data for internal consistency. Examples of consistency checks are given by the following: a male patient who is pregnant; a 50 kg person losing 20 kg in weight; a woman complaining of impotence; and a patient's aging by more than one year over a one-year period.

9.2.5 The duties of a trial monitor

It is essential that all multicentre trials appoint a monitor to visit the various centres and ensure that protocol adherence is satisfactory. It has been suggested that patient can also be used as a monitor [86]. It is true that a patient who is familiar with a trial protocol and expecting a prescription, investigation, questionnaire, or appointment on a particular occasion, may well remind the investigator when these procedures are not arranged.

A monitor's report on a trial may be necessary for the acceptance of the trial results by an administrative body. Table 9-4 gives the duties of a monitor. The monitor should interview the investigators; examine the equipment, make unheralded visits during the progress of the trial, and examine the trial documents. The monitor should ask to see the signed consent forms and the

Table 9-4. Duties of a trial monitor

1. To ascertain that the investigators understand and are familiar with the protocol.
2. To determine that the clinical laboratory or other location for the trial activity has the necessary equipment — for example, materials for resuscitation when specified in the protocol.
3. To make unheralded visits to observe the trial in progress and check adherence to the protocol.
4. To examine the trial documents to ensure that the protocol is adhered to.

detailed protocol. He must check the trial documents for completeness and evidence of adherence to the protocol.

9.3 THE REDUCTION IN VARIABILITY ACHIEVED IN A CROSS-OVER TRIAL

In a cross-over trial a patient takes one treatment, then a further treatment or treatments. The patient acts as his own control and the variability of the response to the treatment is correspondingly reduced. In certain instances the reduction in variability may allow a result to be achieved with a fraction of the number of patients (section 10.8.4). The advantages and disadvantages of cross-over trials are discussed in chapter 11.

9.4 CONCLUSIONS

This chapter has described how the precision of repeated measurements can be estimated together with their accuracy in comparison with a known standard. The relationship between bias, repeatability, and validity has also been discussed. The concept of quality control was introduced and the methods to achieve fewer errors considered, a clear protocol and careful training of the observers being the most important. The detection of errors and the duties of a trial monitor were reviewed.

APPENDIX 9.5

9.5.1 Calculation of the precision and accuracy of measurements

Measurements of systolic blood pressure taken with a mercury sphygmomanometer were compared with two prototype instruments. The first Latin square illustrated in figure 11-3 (section 11.5) was utilised to determine the order. The standard machine was A and the prototypes B and C. Three patients were randomised as patients 1, 2, and 3.

	Machine A	B	C
Patient 1	168 153 151	183 168 174	164 168 165
Mean SD	157.3 9.3†	175 7.5	165.7 2.1
Patient 2	149 149 150	172 164 164	169 178 174
Mean SD	149.3 0.6	166.7 4.6	173.7 4.5

Patient 3		123	110	088
		118	109	110
		118	125	110
Mean		119.7	114.7	102.7
SD		2.9	9.0	12.7
Overall	Mean Precision‡	142.1	152.1	147.4
		5.6	7.3	7.9

$$^{\dagger}\text{Standard Deviation (SD)} = \sqrt{\frac{(168 - 157.3)^2 + (157.3 - 153)^2 + (157.3 - 151)^2}{3 - 1}}$$

$$= 9.3$$

$$^{\ddagger}\text{Precision} = \sqrt{\frac{9.3^2 + 0.6^2 + 2.9^2}{3}}$$

$$= 5.6$$

9.5.2 Conclusion

The results for three patients cannot be conclusive but the prototype machines tended to have a higher precision or standard deviation values than the standard machine. Moreover, the accuracy of the new machines was not acceptable as they read systolic pressure too high in two patients and too low in one.

10. HOW MANY SUBJECTS ARE REQUIRED FOR A TRIAL?

The number of subjects required for a controlled trial depends on the major objective defined for the trial; the level of significance that must be achieved if the objective is reached; the confidence with which a negative result should be reported if the objective is not reached; and the variability of the end-point measurements.

10.1 THE MAJOR OBJECTIVE

The major objective will be the difference to be detected between the intervention and control group. This objective is discussed in section 4.2 and may be in absolute units (for example, a reduction in diastolic blood pressure of 8 mm Hg) or a proportional reduction in an event (for instance, a 50 percent reduction in stroke mortality). The difference should be both biologically important and capable of achievement. If the average antihypertensive drug can lower diastolic blood pressure by 15 mm Hg when compared with placebo and preliminary studies on a new drug reveal a similar effect, it would be sensible to set a reduction of 15 mm Hg as the objective in a trial of this new drug and unreasonable to expect a greater reduction. Even a reduction of 8 mm Hg would be biologically important and would constitute an acceptable objective. Similarly a reduction in stroke mortality of 50 percent appears high but has been exceeded in a trial of antihypertensive treatment [37] and therefore may be a reasonable objective.

Usually the investigator will be interested not only in whether the new treatment is better than the control treatment by a given amount, but also in whether the new and control treatments have a similar effect, or indeed if the new treatment is worse than the control treatment by the defined amount.

10.1.1 The traditional objective: to determine if the new treatment is better, the same, or worse than control treatment

The traditional objective determines: (1) whether the effect (A) of a new treatment, is better than the effect (B) of the control treatment by a given amount; (2) whether the new effect A is comparable with B within certain limits; and (3) whether the unexpected occurs and B is better than A by the given amount. For example: Let the effect of a new drug in lowering diastolic blood pressure be a fall in pressure of A mm Hg. Let the effect of control treatment be a fall in pressure of B mm Hg. The investigator is interested in whether:

1. $A - B > 8 \text{ mm Hg}$
2. $A - B < 8 \text{ mm Hg}$ and $> -8 \text{ mm Hg}$
3. $A - B < -8 \text{ mm Hg}$

This is the traditional aim of a trial but Schwartz, Flamant, and Lellouch [117] have pointed out that there are circumstances where the investigator is only interested in (1) (he only wishes to determine when the new treatment is to be preferred and is not interested if the new treatment is the same as or worse than the old). In certain circumstances the investigator may be interested in (1) and (2) but not (3) (that is, he is not interested when the effect B is greater than A).

10.1.2 The investigator is only interested in whether the new treatment is to be preferred

The example given earlier can be expanded if we assume the investigator is using an established drug, with effect B, in the treatment of hypertension and he is familiar with the effects of the drug and experienced in its use. However, he is willing to change to a new drug with effect A if $A - B > 8 \text{ mm Hg}$. If $A - B$ is not greater than 8 mm Hg the investigator wishes to continue with the established drug. The trial result must decide between two strategies: change to the new drug, and do not change. The investigator must be satisfied with this decision as the only outcome of the trial. The advantage will be a substantial reduction in the number of patients required for the trial. If the established treatment is better than the new treatment this fact may not be apparent as the numbers in the trial will not be sufficient to differentiate between the results $A = B$ and $B > A$.

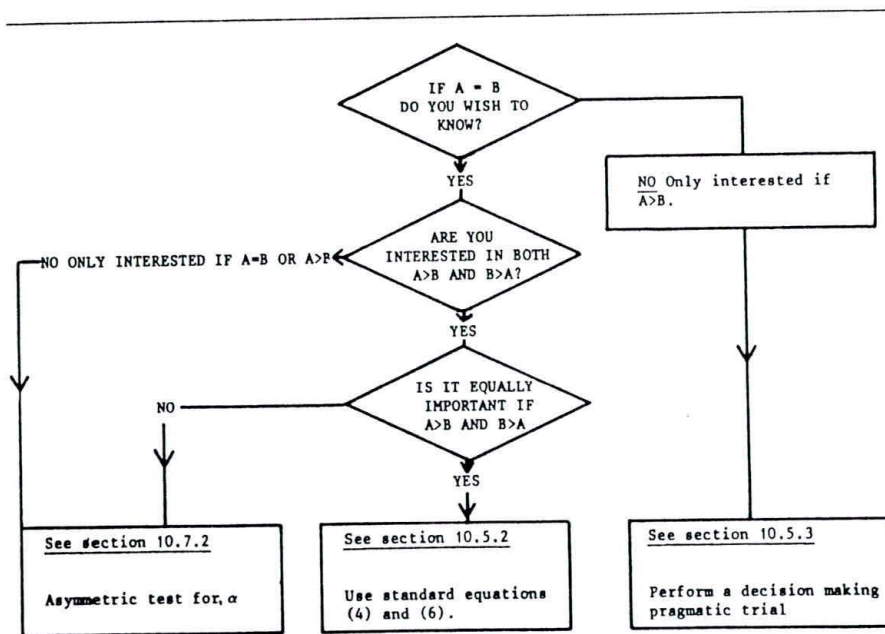


Figure 10-1. Flow chart to determine how to calculate number of subjects for a clinical trial. A is the effect of a new treatment and B the effect of control treatment.

10.1.3 The investigator is certain that the control treatment cannot be better than the new treatment

When the investigator is only deciding whether to change treatment or not, he should not be interested in whether the new treatment is the same as the control treatment or worse. However, when the investigator is interested both in whether A is greater than B and in whether A is similar to B, he must also consider the possibility that the effect B is greater than A. If it is not possible for B to be greater than A, then this concept can be ignored with a reduction in the numbers required for the trial. In biological research, however, it is usually possible for A to be greater than B or B to be greater than A. An investigator using the new blood-pressure-lowering drug would certainly wish to report the fact that the new drug was worse than the old. He should not calculate the numbers required for the trial without taking account of this possibility.

Figure 10-1 provides a flow chart that may help an investigator decide how to calculate the numbers for a trial. The investigator must decide whether he wishes to make a yes/no decision or to quantify the differences between the treatment effects, A and B. If A is the treatment under trial and B the control treatment, the trial organiser must decide whether he is interested in the possibility that $B > A$ and if he is, whether he is equally interested if B is $> A$ as he is in the possibility that $A > B$. If he is not equally interested he may calculate

the numbers using a two-sided asymmetric test (see section 10.7). The figure finally indicates the equations that may be employed to calculate the numbers required for the trial (section 10.5). The central lines of the flow chart indicate the pathway that is applicable to most trials.

10.2 THE STATISTICAL SIGNIFICANCE OF THE DIFFERENCE TO BE DETECTED

The probability that an observed difference is due to the vagaries of chance is measured by a significance test. If the test gives a result less than five percent or one percent the observed difference would be expected to have happened by chance with a frequency of only one in 20 and 100 times respectively. The five percent or one percent levels are arbitrarily cut-off levels at which we decide that the results are unlikely to have happened by chance and the result is statistically significant. These levels must be chosen during the design of the trial (so that the numbers required for the trial can be calculated) and constitute the first type of error, type 1 or α error. More exactly, when the null hypothesis is true (that is, the treatment effects A and B are the same), the type 1 error is the probability of rejecting the null hypothesis when it is true. When calculating the numbers required for a trial, we assume that the desired effect is obtained and $A - B = \delta$ at a level of significance p where δ and p are defined in the major objective.

Figure 10-2 provides a plot representing the frequency of all possible results when the null hypothesis is true and $A = B$. The height of the graph represents the probability of a result, $A - B = 0$ being the most likely, and the area under the graph is set equal to 100 percent. The right-hand and left-hand shaded areas represent the probability that a result will be obtained equal to or greater than $A - B = \delta$ and $B - A = \delta$ respectively. The numbers for the trial are calculated so that, if p must equal five percent and B cannot be greater than A, the right-hand shaded area should represent five percent of the total area. However, in biological experimentation, B can be greater than A and to achieve a two-tailed probability of five percent, each shaded area must equal two and one-half percent as in the figure.

10.3 THE POWER OF THE TRIAL

A negative result for a trial may be difficult to interpret. We must ask whether the number of subjects in the trial is sufficient to demonstrate an important difference between treatments should that difference truly be present. In many instances the numbers of subjects in a trial are so few that a negative conclusion is almost inevitable. The confidence with which we report a negative result is known as the power of the significance test employed. Let us assume that 20 diabetic patients are treated for three months with an active antiplatelet drug and another 20 with placebo for the same period. It is likely that we will not observe a coronary event in either group and if we conclude that treatment is

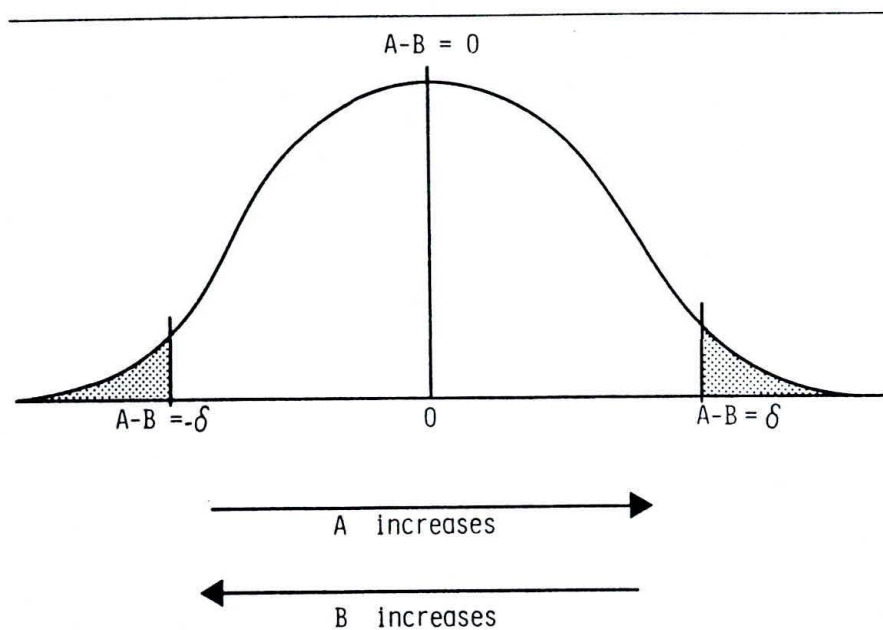


Figure 10-2. The graph represents the expected frequency of results when the effect of one treatment (A) equals the effect of a second treatment (B); that is, the null hypothesis is true. The shaded areas each equal two and one-half percent of the total area under the curve and a result greater than $A - B = \delta$ or less than $A - B = -\delta$ will differ from the expected zero result at the five percent level of significance.

not required for the prevention of ischaemic heart disease it would be obvious that our conclusion is lacking in power.

The error in concluding that a given difference is not present when in reality it is, is known as the type II or β error. The power of trial is the probability of avoiding β error or $1 - \beta$. In figure 10-3a two frequency distributions are drawn: one for the null hypothesis $A - B = 0$ on the left and one on the right for $A - B = \delta$. The numbers for the trial must be calculated so that the probability of a given result appearing compatible with the null hypothesis is small when a true difference of the specified size exists. In the figure the result R lies just within the results expected with the null hypothesis. In this example the numbers for the trial are sufficient so that the distribution for the main objective $A - B = \delta$ does not overlap the distribution of results for the null hypothesis to a great extent. The lined area equals 10 percent and represents the type II error or the probability of a false negative result. The power of the trial is said to be 90 percent. In figure 10-3b only small numbers of patients are entered in the trial and the two distributions overlap each other to a large extent. Again the lined area represents the type II error and is 80 percent, the power of the trial being only 20 percent. For those not familiar with frequency distributions figures 10-2, 10-3a, and 10-3b are represented in figure 10-4 as

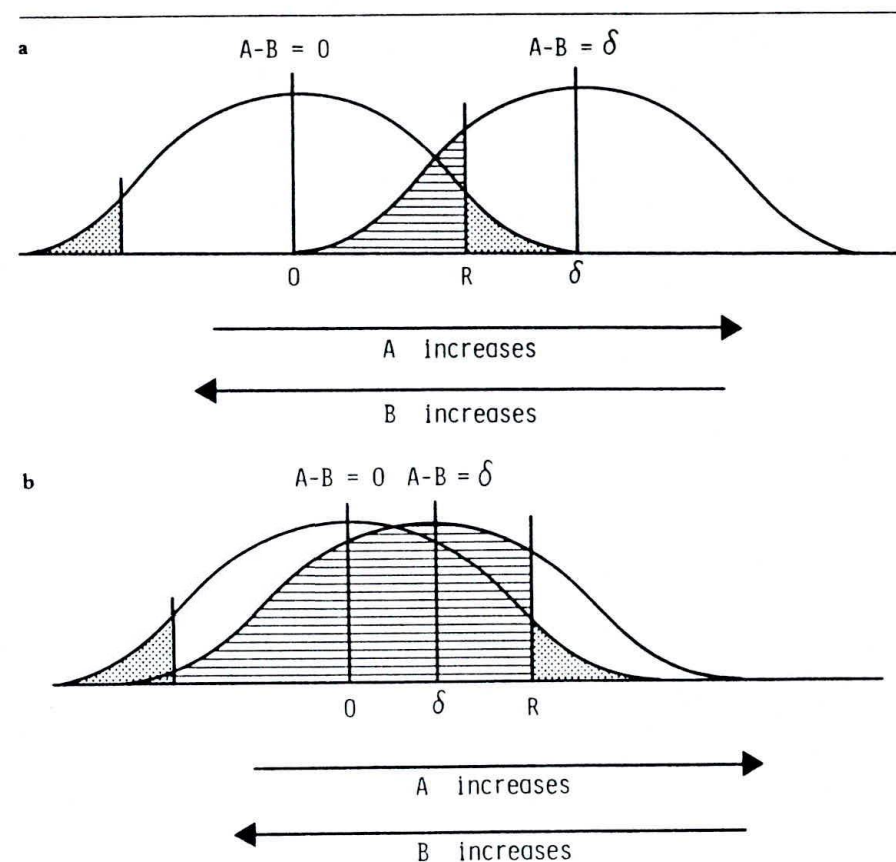


Figure 10-3. Two frequency distributions are represented when $A - B$ is truly equal to zero and when $A - B$ actually equals some difference to be detected, δ . Figure a represents the situation in which there are sufficient numbers of subjects in the trial to ensure that a result R (just compatible with the null hypothesis) has a type II error of 10 percent. \cdot type I error. \equiv type II error. Figure b represents the situation in which there are so few patients in the trial that a result R (just compatible with the null hypothesis) has a type II error of 80 percent. \cdot type I error. \equiv type II error.

10-4a, 10-4b, 10-4c respectively. The area enclosed by each circle equals 100 percent but equal distances around the circles do not correspond to equal intervals of $A - B$. The first circle (10-4a) represents the results when the null hypothesis is true and the dotted areas represent the two-tailed type I error of five percent. Figures 10-4b and 10-4c represent the results when in truth $A - B = \delta$. Figure 10-4b represents the situation when the numbers for a trial are adequate and power is equal to 90 percent and figure 10-4c gives the corresponding representation when the power of a trial is inadequate and equals only 20 percent.

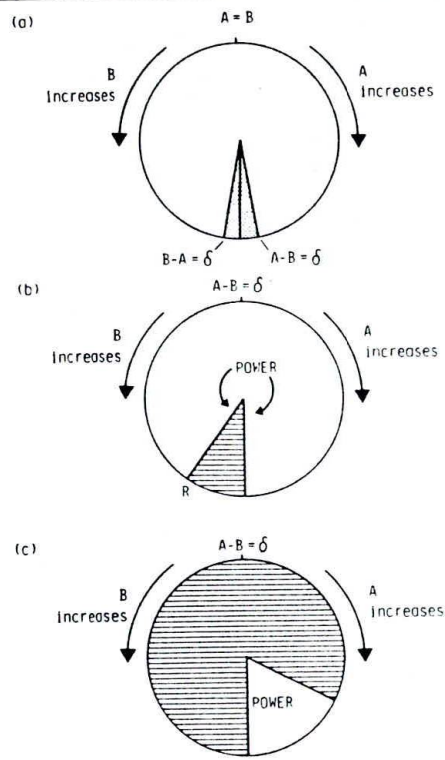


Figure 10-4. Schematic representation of type I error (dotted) and type II error (shaded) for a trial to detect a difference between two treatments A and B. **Figure 10-4a.** The null hypothesis is true, $A - B = 0$, and the probability of two-tailed type I error is given by the dotted area (five percent). **Figure 10-4b.** The null hypothesis is not true and $A - B = \delta$. The trial numbers are sufficient to detect this difference with a power of 90 percent. **Figure 10-4c.** The null hypothesis is not true but the trial numbers are too small and power is only equal to 20 percent.

10.3.1 What power do we require in a trial?

In a trial designed to make the decision between two treatments and not to determine whether the treatments effects are similar, a negative result will not be reported and therefore the power of the trial need not be considered as the trial result will only be "use one treatment" or "use the other treatment." However, with the design and analysis of the standard trial we shall have to decide on the power we require.

Table 10-1 gives the results of antihypertensive treatment in four placebo-controlled trials. The first four rows give the effect of treatment on stroke incidence. In three trials in male patients treatment reduced stroke incidence and this was statistically significant in two trials. There was no effect of treat-

Table 10-1. The results of four trials of active antihypertensive treatment versus placebo.

Table 10-1. The results of four trials of active antihypertensive treatment versus placebo.								
Trial	Annual rate/1000		Significance of difference ($P_1 - P_2$)	The difference ($P_1 - P_2$) that would be detected at 5% level	Power of trial given ($P_1 - P_2$) significant at 5% level (%)	Number in each group required to reach 5% significance, 90% power	Length of trial (years)	Actual number in each group
	Control group (P_1)	Treatment group (P_2)						
STROKE								
Veterans Administration [47]	33	8	$P = 0.30$	36	23	360	2	72
Veterans Administration [37]	31	10	$P < 0.01$	16	75	340	3	190
Hamilton [118]	83	0	$P < 0.05$	79	55	35	4	11
Males	39	38	$P = 0.98$	43	3	56,000	4	20
Females								
MYOCARDIAL INFARCTION								
Veterans Administration [47]	33	0	$P = 0.10$	36	42	160	2	72
Veterans Administration [37]	20	18	$P = 0.90$	13	4	43,000	3	190
Hamilton [118]	21	0	$P = 0.40$	31	17	140	4	11
Males	26	13	$P = 0.60$	33	9	570	4	20
Females								
Cooperative randomised controlled trial [98]	21	0	$P = 0.40$	31	17	280	2	25
Males	16	13	$P = 0.90$	25	3	15,000	2	33
Females								

Note: The effect of treatment on stroke and myocardial infarction rates is given together with the power of the trials, the number of patients who were in the trials, and the number who would be required to demonstrate the differences at the 5% level of significance and 90% power.

ment in women. One trial [98] did not observe a stroke event in either the actively treated or the control group; this trial was excluded from this section.

The final six rows of the table consider the effects of treatment on the incidence of myocardial infarction. There was not a significant reduction in myocardial infarction in any single trial, but in males the reductions were 100 percent, 10 percent, 100 percent, and 100 percent respectively. The fourth column of the table gives the difference between the treatment and control groups which would have been statistically significant at the five percent level. The fifth column gives the power of the trial, given that the actually observed difference was significant at the five percent level (calculated as in section 10.6). The sixth column gives the numbers in the trial that would be needed to prove with 90 percent power that the observed difference was significant at the five percent level and the last column gives the actual numbers in the trials. The following points may be made:

1. In the three trials showing a large effect of treatment on the incidence of myocardial infarction (47, 118, 98), the numbers entered into the trial were only 45 percent, eight percent, and nine percent of the numbers required.
2. The power of these three trials was 42 percent, 17 percent, and 17 percent respectively.
3. The power of the trials that successfully demonstrated a reduction in stroke incidence (37, 118) was 75 percent and 55 percent respectively.

These calculations help us to get a feel for the power term that we might include in the calculation of the numbers required for a trial. Unlike the level of significance (type I error), there is no time-honoured tradition stating what is acceptable for type II or β error. A low-power term of 50 percent has been employed [119] as has a high level of 95 percent [120, 121].

10.3.2 Confidence limits rather than power should be reported for negative results

The confidence limits for the result of a trial give the range of figures that are compatible with the observed result for a given probability. Section 15.6 shows how to calculate these limits as these are more easily understood than power terms.

10.3.3 A clear demonstration of the power of trials

Rose insisted that confidence limits should be reported for the results of any negative trial [122] and had in mind the use of 95 percent confidence limits, corresponding to the usual five percent test of significance. Baber and Lewis took up his point and, calculating less stringent 90 percent confidence limits, they showed that in 18 trials of the use of beta-blockers following myocardial infarction, the confidence limits encompassed a 50 percent reduction in mortality in 14 and an increased mortality (of any degree) in 16 (123).

10.4 VARIABILITY OF THE RESULTS

The number of subjects required for a trial will depend on the variability of the end points being measured. Either quantitative or qualitative measurements may have to be considered and their variability calculated.

10.4.1 Quantitative data

When the trial is designed to detect changes in quantitative information, such as weight, blood sugar, or blood pressure, the variance of the measurement must be calculated. In a between-patient trial (two separate groups), the variance of the observation itself has usually been determined in preliminary studies and is calculated from the following formula:

$$\text{Variance} = \frac{\sum (x - \bar{x})^2}{n_a - 1} \dots \dots \dots (10.1)$$

Where x is one of a number (n_a) of available and normally distributed observations (one per subject), \bar{x} is the mean observation and $\sum (x - \bar{x})^2$ denotes the sum of the squared differences between individual observations and the mean. Occasionally this end point of a trial may be an average of more than one measurement on an individual. For example, if three measurements of weight are averaged as an end point for a trial, the averages will be entered as the x 's and the number, n_a , will equal the number of subjects.

For the within-patient cross-over trial or the change from baseline in a between-subject study, the variance of the change in the measurement must be calculated:

$$\text{Variance of change} = \frac{\sum (d - \bar{d})^2}{n_a - 1} \dots \dots \dots (10.2)$$

Where the notation is as for equation (10.1) except that d and \bar{d} refer to the difference between two observations and not the observations themselves. The variance of a change may be considerably less than the variance of the original observations, thus reducing the numbers required for a cross-over trial (section 10.8.4).

10.4.2 Qualitative or discrete data

Qualitative data include measurements such as the proportion of patients who die, improve, or relapse. The variance of such data depends on the proportion expected in the control and treatment groups. With quantitative data the variance of data in the control and treated groups is expected to be the same, but in qualitative data the variance will depend on the expected values.

$$\text{Variance in control group} = \frac{P_1 (1 - P_1)}{n} \dots \dots \dots (10.3)$$

MP-100
14148

where P_1 is the proportion with the end point in the control group and n the number of subjects that will be required in the control group.

Similarly:

$$\text{Variance in the intervention group} = \frac{P_2(1 - P_2)}{n}$$

where P_2 is the proportion with the end point in the intervention group and the same number of subjects (n) is assumed. To calculate the numbers for a trial with a qualitative end point the expected proportion with the end point in the control group (P_1) must be known from existing data; P_2 can then be calculated from the trial objective. For example, if the major objective is to reduce mortality by 50 percent, $P_2 = 0.5P_1$.

With qualitative data, P_1 and P_2 are required to calculate the number required for the trial but the number of available subjects (n_a) required for the original observation of P_1 is not required.

10.5 CALCULATING THE NUMBERS

Some useful formulae will be presented but their derivation is given in the statistical texts that are referenced. We must first consider whether the investigator wishes to estimate the effect of the new treatment or to make a decision whether or not to use it (or investigate it).

10.5.1 The usual or explanatory trial

Schwartz and his colleagues considered the usual trial to be an attempt to examine the magnitude of treatment effects and to explain the observations [117, 124]. In figure 10-3a the classical situation is illustrated when the control and intervention group results are truly identical, giving a difference in results of zero, and when they are truly different by a difference δ . The results of the trial will be distributed around mean zero in the first instance and around mean δ in the second instance giving the two frequency distributions. The number of patients required for the trial must be arranged so that the difference designated as the objective of the trial (δ) is associated with a given.

10.5.1.1 Type I error, α (two-tailed test)

In figure 10-2, δ is arranged to define two-and-one-half percent of the area under the right-hand curve, and the distance 0 to δ is said to be 1.96 standardised normal deviates.

To calculate the numbers for a trial the number of standardised normal deviates has to be provided for a given α . Statistical textbooks provide tables of these normal deviates.

10.5.1.2 Type II error, β (power = $1 - \beta$)

In figure 10-3a if a just-significant result R or less is arranged to occupy ten percent of the area under the right-hand curve (power 90 percent), then the corresponding standardised normal deviate (distance OR) will be 1.28.

10.5.1.3 Type III error, γ

A type III error occurs when we conclude that the truly better treatment is actually the worse treatment. In the explanatory trial type III error is vanishingly small.

10.5.2 Calculating the numbers for the classical explanatory trial

It will be assumed that either treatment may be superior giving a two-tailed test for α .

10.5.2.1 Quantitative data

Let n be the minimum number required in each of two groups.

Let d be the difference to be detected.

Let K be a constant equal to the square of the sum of the standardised normal deviates for α and β ($ND_\alpha + ND_\beta$)²

Let s be the standard deviation of the measurement, then

$$n = \frac{2Ks^2}{d^2} \dots \dots \dots (10.4)$$

when $\alpha = 0.05$ and $\beta = 0.10$, $K = (1.96 + 1.28)^2 = 10.5$ and α is two-tailed and β one-tailed.

10.5.2.2 Qualitative data

Let P_1 be the proportion having the event in the control group.

Let P_2 be the proportion in the intervention group.

Let n and K be as in section 10.5.2.1.

$$n = \frac{K[P_1(1 - P_1) + P_2(1 - P_2)]}{(P_1 - P_2)^2} \dots \dots \dots (10.5)$$

This is an approximate formula and a more exact formula is [125]:

$$n = \frac{(ND_\alpha \sqrt{2\bar{P}(1 - \bar{P})} + ND_\beta \sqrt{P_1(1 - P_1) + P_2(1 - P_2)})^2}{(P_1 - P_2)^2} \dots \dots \dots (10.6)$$

Where $\bar{P} = \frac{1}{2}(P_1 + P_2)$

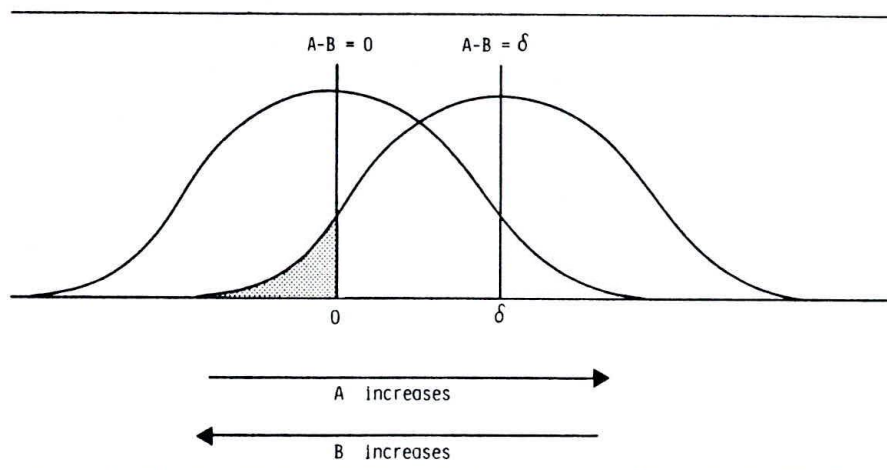


Figure 10-5. Two frequency distributions are represented where $A - B = 0$ (the left curve) and $A - B = \delta$ (the right-hand curve). The investigator is not interested in whether A and B are equal but only wishes to decide between the use of treatment A and treatment B. Type I error is 100 percent, type II error is zero, and the type III (γ) error is given by the dotted area. This area equals five percent and gives the probability of choosing the worst treatment: B, when A is the best.

The proportion of events in the experimental group can be adjusted for the number of dropouts to be expected from default and from withdrawal due to death from causes other than trial end points. Also, the numbers can be adjusted if the treatment takes a time to achieve full benefit [121]. Similarly the formulae may be adapted for more than two drug groups and for unequal allocation of patients between the groups (section 10.11).

George and Desu [126] have also discussed the situation where survival times rather than events are to be compared.

10.5.3 The pragmatic decision-making trial

Schwartz and Léllouch named a decision-making trial a pragmatic trial [124]. Figure 10-5 illustrates the situation when we wish to make the decision between treatments A and B but do not care if they are truly similar. δ is the mean result for a distribution when a true difference is present and 0 the mean when there is no difference between the treatments. We must note the following:

1. The type I error (α), or the likelihood of saying one treatment is to be preferred when the treatments are equal is 100 percent.
2. The type II error (β), or likelihood of saying that the treatments are equal when they are not, is zero.
3. The type III error (γ), the probability of preferring the worst treatment can be large and the trial numbers must be arranged to limit this error. If the

Table 10-2. The three kinds of errors. Treatment effects A and B. In reality the treatment effects may be equal ($A - B = 0$) or unequal $B > A$ or $A > B$.

		Trial conclusion		
		$B > A$	$A - B = 0$	$A > B$
T	$B > A$	—	β	γ
R				
U	$A - B = 0$	$\frac{1}{2} \alpha$	—	$\frac{1}{2} \alpha$
T				
H	$A > B$	γ	β	—

α is the type I error; β the type II; and γ the type III error [117].

error is limited to five percent the distance, 0δ , will be 1.64 standardized normal deviates.

Table 10-2 illustrates the three kinds of error according to the observed results and the true results.

The calculation of the numbers required for a pragmatic trial is performed using formulae (10.4) and (10.5) in section 10.5.2 but K now equals the square of the standardised normal deviate for γ or type III error (one-tailed test) [117]. If γ is set at five percent, $K = 1.64^2 = 2.7$.

We can also decide on the use of a new drug when a new drug A is better than an established drug B by a certain number of units (D). The decision to use A could be made for results $> D$ and a decision in favour of B made when the result is less than D . The employment of a difference, D , in making a decision does not affect the numbers required for the trial [117].

10.6 CALCULATION OF THE POWER OF A REPORTED TRIAL

The concept of the power of a significance test was discussed in section 10.3. The power of a reported trial when the observed difference is assumed significant at the five percent level is given by ND_{β} when calculated by substitution of n and ($ND_{\alpha} = 1.96$) in K and equations (10.4) and (10.5). This calculation of power led to the results reported in table 10-1 (section 10.3.1).

10.7 HOW THE SPECIFIED LEVELS OF ERROR INFLUENCE THE NUMBER OF PATIENTS REQUIRED FOR A TRIAL

The level or presence of any error term employed in calculating the number of subjects will be influenced by whether the trial is required for estimation of effect (explanatory trial, section 10.5.1) or as a mechanism for decision-making (section 10.5.3); by whether the α error is symmetrically two-tailed or not; and by the effect on α error of taking repeated looks at the data. In addition we must examine the effect of factors influencing the level of variance of the data (section 10.8); the treatment effect that is to be detected (section

Table 10-3. Increase in numbers required for a trial with either $\alpha = 1\%$ or 5% and for increasing levels of power. (α , two-tailed.)

$1 - \beta$	Increase in Numbers Required for a Trial				
	50%	75%	90%	95%	99%
$\alpha = 5\%$	1	1.8	2.8	3.4	4.8
$\alpha = 1\%$	1.7	2.8	3.9	4.6	6.3

10.9); the effect of dropout and withdrawal (section 10.10); and whether or not we require equal numbers in the control and treatment groups (section 10.11).

10.7.1 The level of type I (α) and type II (β) error

Table 10-3 gives the increase in trial numbers that would be expected if α equals one percent and not five percent, and if $1 - \beta$ (power) equals 75 percent, 90 percent, 95 percent, and 99 percent rather than 50 percent. When α equals one percent and power 90 percent the trial numbers are increased fourfold and with a power of 99 percent the numbers are increased over sixfold.

10.7.2 Asymmetric instead of a two-tailed test for α

Schwartz and his colleagues have argued the case for the one-tailed test of significance [117] where the possibility that the intervention treatment is worse than the control treatment is ignored. However, we are usually equally interested in whether an intervention treatment is better or worse than a control treatment and a one-tailed test is not applicable. Schwartz and his co-authors pursued the argument and suggested that we may be interested in an unexpected finding but only if it is highly significant. They provided an example where the research worker was interested in a right-sided $2\frac{1}{2}$ percent probability as in figure 10-2 but only in a 0.1 percent probability in the other direction. In their example the power of detecting a significant ($P < 0.025$) difference in one direction was 95 percent whereas the power of detecting a difference in the other direction was 70 percent. They stated, "Thus we accept a rather large probability of failing to detect a difference on the left. This is quite reasonable; the test is primarily intended to detect a difference on the right . . . A significant result on the left has to be considered as a byproduct."

A by-product it may be, but although a trial of a treatment is usually mounted to detect a benefit, and an adverse effect of treatment is not anticipated (it would be unethical to perform the trial if it were), we must agree that the failure to detect an adverse effect may have the most serious consequences. If a trial lacks power in detecting a benefit, then patients who could be helped may not receive the treatment and this will be to their detriment. However, most trials are performed when a clinical impression suggests that the treatment is beneficial and one trial result showing no difference from, say placebo, will often result in the continued use of the drug. But a single trial showing a

significant adverse effect can result in the treatment's being withdrawn from use, used less frequently, or limited to certain patients. It can be argued that the detection of an adverse effect has more influence on patient care and is more important than the demonstration of a beneficial effect. It is only rarely that the investigator can justify an asymmetric test for type I error.

10.7.3 The effect of repeated looks at the data on the type I error

A statistical test, significant at the five percent level, indicates that the observed result has less than a five in 100 likelihood of having occurred by chance. However, the assumption is made that the statistical test is only performed once during the course of the trial. If the investigator makes the test after 20 patients have entered the trial, then after 40, 60, 80, etcetera, he will greatly increase the odds on reaching a five in 100 chance. Moreover he will presumably stop recruitment and end the trial when a significant result is observed. McPherson has pointed out that ten repeated tests on accumulating data at the one percent level of significance during a trial will be the same as an overall test for the trial at the five percent level of significance [44]. Similarly ten tests at the five percent level of significance will lead to an overall significance level test of 19 percent, almost a one in five probability of the finding being due to chance. In a long-term trial, repeated testing may be necessary for ethical reasons. Table 10-4 gives the significance level for individual repeated tests (nominal levels of significance), one of which would have to be exceeded in order to achieve an overall level of significance of one percent or five percent. The maximum number of repeated assessments has to be decided at the planning stage of the trial and the nominal level of significance selected at that stage to give an overall level of significance of one percent or five percent.

The nominal level of significance must be strictly adhered to. For example, with a maximum of ten tests, a nominal level of significance of one percent, and an overall level of five percent, the first test may be significant at the three percent level, but the trial cannot be stopped at this stage. The first statistical test is one of a series of ten and is as likely to be falsely significant as the last. Even though only one test has been performed and that was significant at the three percent level an overall level of significance of five percent cannot be claimed. The decision rule was to stop the trial if the one percent level is

Table 10-4. Nominal levels of significance (%) that must be exceeded if an overall level of significance is to be achieved of 1% or 5% when the test is repeated 5, 10, 15, or 20 times [44].

		Number of Repeated Tests			
		5	10	15	20
True level of significance (%)	1	0.28	0.19	0.15	0.13
	5	1.59	1.07	0.86	0.75

achieved in any one of ten tests and only if this level is reached in the first test may the trial be stopped.

If the investigator wishes to review the data constantly, perhaps with a view to an early completion of the trial, he may adopt a sequential trial design. This design is discussed in section 11.7 and takes into account the use of repeated significance tests.

10.8 HOW THE VARIANCE OF THE DATA INFLUENCES THE NUMBERS REQUIRED

10.8.1 Quantitative data

With quantitative data the variance of a result may be reduced by replication of the measurement. For example, it is reasonable to assume that the variance of fasting blood glucose will be reduced if it is measured every day for three days and the average taken as the end point of the trial, rather than if a single result is employed. The variance of an average result for each patient will be lower than that of a single measurement.

10.8.2 Qualitative data

With proportional data the variance is maximum at 0.50 and minimum at very high and very low values. More importantly, however, the higher the rate in the control group the greater the effect of treatment for a given proportional reduction. For example, a 50 percent reduction in a 90 percent event rate gives a fall in events to 45 percent (a difference of 45 percent) whereas a 50 percent reduction in a 20 percent event rate gives a fall to only 10 percent (a difference of 10 percent). The higher the proportion in the control group getting an event, the lower will be the numbers required for the trial. Intuitively, the more patients that experience an event in the control group, the fewer will be the numbers required and this is one of the reasons for doing trials in selected high risk patients. However, Sondik and colleagues [127] considered a trial in which subjects with a high serum cholesterol were entered. The higher the serum cholesterol, the higher the risk. But if the subjects are to be detected by screening, the greater the serum cholesterol required to enter the trial, the larger the number of subjects that have to be screened. In such circumstances it may be less expensive overall to enter medium-risk patients to a trial.

10.8.3 Reduction in variance by measuring a change from baseline

In certain circumstances, measuring change from baseline can drastically reduce the numbers required compared with a trial without a baseline. Table 10-5 gives the numbers required for a trial of a new antihypertensive drug. The standard deviation of between-subject blood pressure measurements (13 mm Hg) was derived from a trial of such a drug and is given in column one. Similarly the standard deviation of within-patient changes in blood pressure is given in columns two and three (8 mm Hg). The difference to be detected is 10 mm Hg, α = five percent (two-tailed), β = ten percent. In a between-patient trial comparing two treatments without a baseline measurement equation

Table 10-5. Numbers required for a trial to show that one antihypertensive drug reduces diastolic blood pressure by 10 mm Hg more than a second drug.

	Between-Patient Study		Within-Patient Study
	Absolute values	Change in pressure	Change in pressure
Difference to be detected (mm Hg)	10	10	10
Standard deviation (s and s_w)	13	8	8
Significance level	5%	5%	5%
Power	90%	90%	90%
K	10.5	10.5	10.5
Number of patients to be recruited to the trial	70	27	7
Number of observations required	70	54	14

Note: A between-patient design with and without baseline measurements is compared with a within-patient (cross-over) trial.

(10.4) estimates $n = 35$ so that 70 patients with 70 observations are required for the trial. With the same criteria but estimating change from baseline only 27 patients and 54 observations are required.

10.8.4 Reduction in numbers using the patient as his own control (a cross-over trial)

With the same parameters but a cross-over (within-patient) trial the formula is:

$$n_w = \frac{KS_w^2}{d^2} \dots \dots \dots (10.7)$$

where n_w is the number of subjects, $2n_w$ the number of observations, and S_w the standard deviation calculated from within-patient changes in blood pressure. In this example only seven patients giving 14 observations would be required for the trial but we have assumed that the difference between the treatments is the same in the first as the second treatment period. Cross-over (within-patient) trials are discussed further in section 11.2.

10.9 THE EFFECT OF ALTERING THE DIFFERENCES TO BE DETECTED BETWEEN CONTROL AND TREATED GROUPS

The smaller the difference to be detected the larger the numbers required for the trial. Table 10-6 gives the number of patients who might be required in a trial of a treatment to reduce the frequency of reinfarction in patients who have already suffered one myocardial infarction. It is assumed that the type I error is five percent and the power term 90 percent. The numbers are provided accord-

Table 10-6. Number of patients required for a trial of secondary prevention in myocardial infarction, according to the event rate in the placebo group over the duration of the trial and the % reduction to be determined in the intervention group.

Event rate in placebo group/100	% reduction in events in the intervention group	Total number required for the trial
10	10%	29,460
10	20%	7,010
10	30%	2,960
10	40%	1,570
10	50%	950
20	10%	13,180
20	20%	3,160
20	30%	1,340
20	40%	720
20	50%	430

Type I error = 5%; power = 90%.

ing to the rate in the placebo group over the duration of the trial and the expected reduction in events in the treated group. In this example 30 times as many patients are required to detect a ten percent reduction compared with a 50 percent reduction [121].

10.10 THE EFFECTS OF DROPOUT ON THE NUMBERS REQUIRED

Dropout is discussed in section 14.3 and is important if withdrawn subjects are to be excluded from the analysis. If they are retained in the analysis they may dilute any effect of treatment. Dropouts consist of those who default or otherwise do not follow the trial protocol; those who are withdrawn for criteria unrelated to the end point of the trial; those who die from causes unrelated to the trial end points; and those who are withdrawn from the trial for criteria possibly related to trial end points.

The Biometrics Research Branch, National Heart Institute [121] have published tables giving the numbers required for trials according to the expected dropout rate; these tables are appropriate when such subjects are to be excluded from the analyses.

10.11 DO WE NEED EQUAL CONTROL AND TREATMENT GROUPS?

The concept of unequal randomisation has been introduced in section 7.3. Traditionally, treatment allocation has been arranged so that each group contains an equal number of subjects. However, two alternatives have been suggested, the use of 2:1 allocation of new:old treatments in trials comparing two treatments [70] and a relative increase in the number in the control group [120] when several treatments have to be compared with this group.

10.11.1 A 2:1 allocation ratio with two treatment groups

A 2:1 allocation has been suggested for a comparison of a new treatment with an old or placebo treatment [70]. Less will be known about the effects of the

new treatment and this is one reason for increasing the numbers receiving that treatment. A new treatment may also have to be compared in two different modes of administration or dose schedules, and one dose may be given to one-third of subjects, the second dose to another third, and the old control treatment to the remainder, thus resulting in a 2:1 allocation. Peto and his colleagues [70] suggest that this strategy may allow different groups to participate in a trial of a new treatment even when they have divergent views on minor variants of treatment. Unequal allocation gives some loss of efficiency compared to a 1:1 allocation, but a 2:1 allocation is equivalent to performing a 1:1 allocation and eliminating about 10 percent of the patients from the trial. However, more unequal comparisons cannot be supported and 3:1 randomisation is equivalent to eliminating a quarter of the patients from the trial.

Unequal randomisation may also be employed when the costs of treatment vary. Cochran [128] and Nam [129] have discussed the square-root rule which states, "If it costs r times as much to study a subject on treatment A than B then one should allocate \sqrt{r} times as many patients to B than A." This procedure minimises the cost of a trial while preserving power. Gail and colleagues [130] considered a similar situation where one treatment was more hazardous than the other and developed a case-saving rule.

10.11.2 Unequal randomisation when comparing more than two groups

When more than one treatment group is to be compared with a standard control treatment, it may be desirable to increase the relative number receiving the control treatment. For example, the Coronary Drug Project trial compared five treatment groups with a placebo group. As each of the five drug groups was to be compared with the same placebo group, it was necessary to determine the final mortality in the placebo group with greater precision than in the actively treated groups (it was expected that the five-year mortality in the placebo group would be 30 percent and that active treatment would reduce this rate by a quarter). The Coronary Drug Project allocated 2.5 times as many patients to the placebo group (2,793 patients) as to any individual actively treated group (1,117 patients in each group) [120]. The ratio 2.5 : 1 was calculated by minimising the variance of the difference between the results in the drug groups and in the placebo group [131].

10.12 AIDS TO CALCULATING THE NUMBERS REQUIRED FOR A TRIAL

Tables have been published of the numbers required for a trial given the dropout rate, the difference to be detected, the event rate in the control group, and the duration of the trial [121]. As an example, figure 10-6 gives a graphical representation of the numbers required in each group according to the percentage of patients expected to respond to two treatments, α = five percent, $1 - \beta$ = 50 percent [119]. If, for example, 60 percent respond to one treatment and 40 percent to the other, between 40 and 50 patients will be required in each group.

NUMBER OF PATIENTS REQUIRED FOR CLINICAL TRIALS
50% CHANCE OF SUCCESS 5% LEVEL OF SIGNIFICANCE

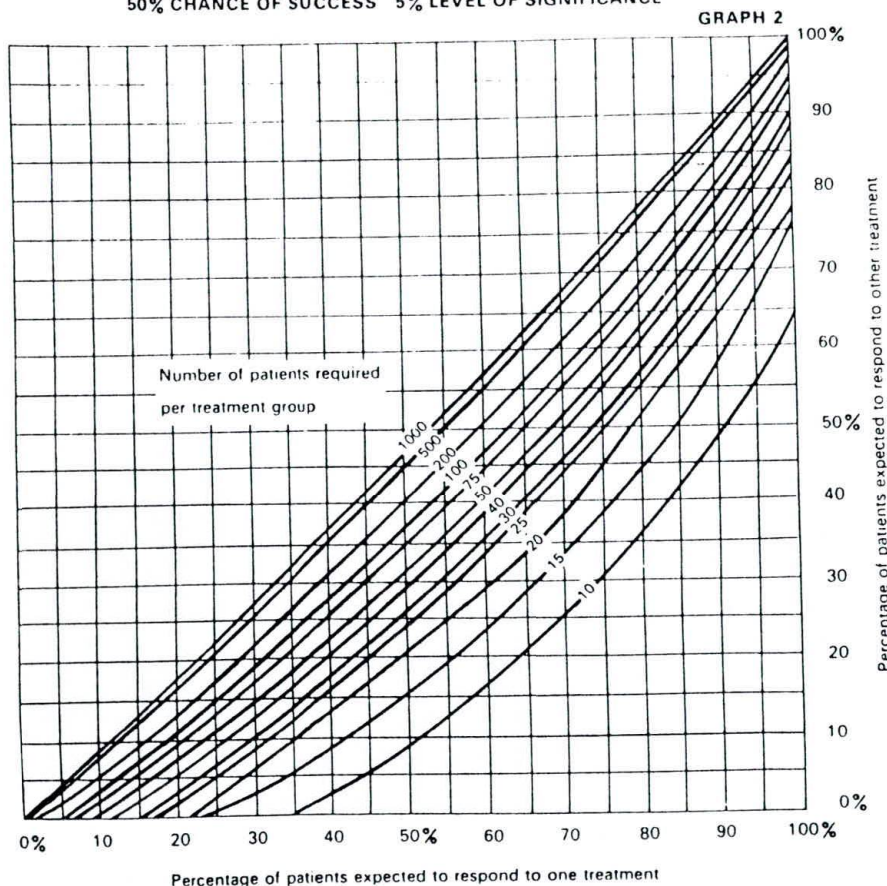


Figure 10-6. Number of patients required for a randomised controlled trial: 50 percent power, five percent type I error. Reproduced with permission from Clark, C.J., and Downie, C.C. *Lancet* 2: 1357, 1966.

10.13 FAILURE TO PREDICT THE VARIANCE OF MEASUREMENTS IN THE TRIAL OR THE FREQUENCY OF EVENTS IN THE CONTROL GROUP

Calculation of the number of patients required for a trial requires an accurate estimate either of the variance of measurements used as the trial end point or of the frequency of events in the control group.

estimate of the variance of measurements

If the variance of measurements during the trial is lower than estimated, then the number of patients required is less than calculated; if the variance is larger, more patients are required. Either situation may occur: the initial estimate of

variance may have been low if determined under perfect standardised conditions that are not reproduced in the trial; on the other hand, the variance may have been high if calculated from observations in normal clinical practice. In the trial, standardisation and close attention to detail may reduce the variability of the results.

10.13.2 An inaccurate estimate of the number of events in the control group

In a clinical trial, careful attention to the patients' welfare may markedly reduce the number of events that were expected in the control group. This effect will be increased if patients are withdrawn from the trial prior to an event owing to an observed deterioration in their condition. In the Hypertension Detection and Follow-up Program trial [132] the patients who were closely followed for hypertension showed a reduction in mortality from several conditions unrelated to hypertension, possibly the effect of an early detection of other disease processes.

It is possible, in a long-term trial, to observe more events in the placebo group than expected. This can result from a failure to allow for the ageing of the population. In a ten-year trial of patients aged, say, 60, the correct number of events should be based on the frequency of events at age 65. If the number of expected events is based on the number at age 60, more events may be observed during the trial than expected owing to this error.

10.14 CONCLUSIONS

In order to calculate the numbers required for a randomised controlled trial, the major objective must be defined exactly. The treatment effect to be detected has to be designated and the distinction made between a trial designed to arrive at such an estimate (an explanatory trial) and one intended only to reach a decision (pragmatic trial). The investigator must define the level of type I and type II errors he will allow and also calculate the variance he expects in the trial end point. The methods of calculating the numbers required for a trial were given in this chapter together with the effect on these numbers of changes in the objective, level of type I and type II error, and change in variance of the trial end point. More specifically, the advantages of the within-patient crossover trial were discussed and the advantages and disadvantages of using unequal allocation procedures were given. The problems of a negative result were considered in some detail and the power of a trial defined and calculated. The concept of confidence limits was introduced and the problem that repeated looks at the data raises in determining the type I error was reviewed. Aids were considered for calculating the numbers required for a trial. Lastly, it was admitted that the assumptions made in the calculations may prove to be in error. However, it is very important to attempt to calculate the numbers required for a trial. Otherwise an inappropriate design and protocol may be adopted, leading to a predictably inconclusive result.

11. DIFFERENT TRIAL DESIGNS

In the standard trial design a subject or patient is considered suitable for entry to the trial and is then randomized to a treatment group; each group receives a single treatment. A fixed number of persons enter the trial and are followed for a predetermined interval of time; treatment is stopped at the end of this period. Other trial designs have to be considered. The subject may be asked to take more than one treatment consecutively (cross-over trial), more than one treatment simultaneously (trial to detect an interaction between treatments), or more than one treatment consecutively and concurrently (cross-over trial to detect an interaction). The remaining design variations do not involve a fixed number of persons being entered. The standard trial, cross-over trial, trial to detect an interaction between treatments, and cross-over trial to detect an interaction are illustrated in figure 11-1 for four treatments A, B, C, and D.

The standard design (number 1) allows four treatments to be compared simultaneously and has been called a parallel-groups trial. If one treatment is a placebo then the effect of the other three treatments can be estimated in the full knowledge of any placebo effect or change in baseline measurements. In the cross-over trial (number 2) each subject receives the four treatments and the order of treatment is randomized. In this example four different orders are specified. Trial design number 3 allows the effect of one treatment to be assessed in the presence of a second treatment (that is, the presence of an interaction between treatments may be determined). The individual drug effects are also determined as a fourth group receives neither drug (but usually

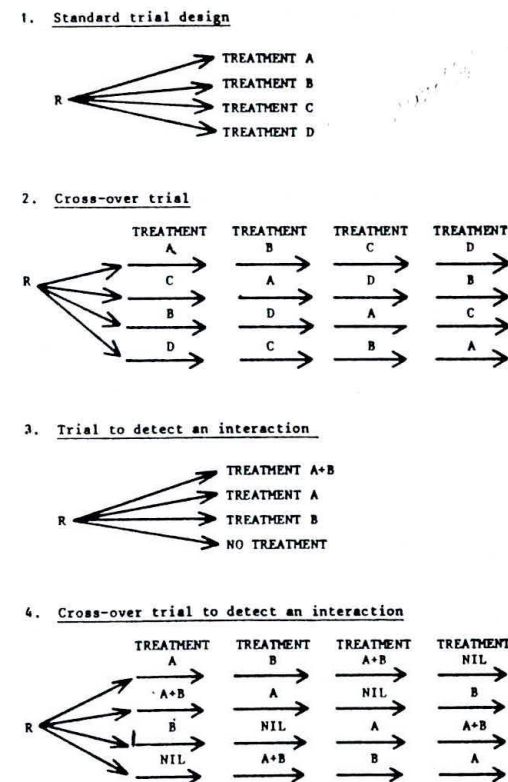


Figure 11–1. Two designs to assess the effects of four treatments A, B, C, and D and two designs to test for an interaction between two treatments A and B. R = point of randomisation.

1. Standard trial design. Four groups studied in parallel to assess four treatments.
2. Cross-over design to assess the effects of four treatments within patients.
3. Parallel groups study to detect an interaction between two treatments.
4. Within-patient cross-over trial to detect an interaction.

receives a placebo). In design 4 an interaction between treatments can be determined within-subject in a cross-over trial.

11.1 THE USE OF THE STANDARD TRIAL DESIGN

The standard design is the one most commonly used and has the virtue of simplicity in that a single treatment is given to each group and a fixed number of patients is involved. The other frequently employed design, the cross-over trial, is inappropriate when a treatment is curative, when the duration of treatment has to be long, when the effects of treatment persist for some time after stopping treatment (a carry-over effect), or when a large number of treatments have to be compared. In any of these circumstances the standard design has to be employed. This design is also more appropriate when a large

number of subjects are available for the trial. The standard design should be employed in the following circumstances.

11.1.1 When treatment is curative

If the trial is to test a curative treatment for an illness, the cross-over design cannot be employed. There is no point in a cured patient continuing with further treatments.

11.1.2 When the duration of treatment has to be long

If the effect of a drug has to be determined after, say, five years, a cross-over trial may take too long since the duration of the trial is ten years with two treatments and 20 years with four.

11.1.3 When the effect of one treatment is different when it follows another treatment

If the effect of one drug persists for a long time, a carry-over effect of this treatment into the next treatment period may interfere with the effect of any further treatment and the standard trial design is to be preferred.

When comparing two treatments in a cross-over trial the difference between the treatment effects must be independent of the order of administration. Hills and Armitage have concluded that if previous experience with the treatments has not proven that this is true, then a parallel group study should be carried out [133]. A different result in one period (in statistical terms, an interaction between treatment and period) may not only occur with a carry-over effect. For example, placebo treatment may possibly be more effective when given first to lower blood pressure or when given last to relieve a painful condition that is improving with time.

11.1.4 When a large number of treatments are to be compared

With a large number of treatments the trial would be too long and complex if each subject had to take every treatment. The standard trial allows several treatments to be compared.

11.1.5 When the number of subjects available for the trial is unlimited

When large numbers of subjects are available for the trial we can assume that enough subjects will be recruited in a standard design to detect important difference between the treatments, even if the variation of between-subject measurements is much greater than the variation of within-subject measurements. Assuming costs are not the limiting factor, we can opt for the less efficient design but one that is quicker and simpler to execute without the difficulty of carry-over effects and because it is shorter fewer patients will drop out.

11.2 THE USE OF THE CROSS-OVER TRIAL DESIGN

The cross-over design can be considered when the condition being investigated is constant and only temporarily affected by treatment. For example, a patient with a high blood pressure or blood sugar may receive drug treatment that has a short-term effect on his condition and may then take a succession of treatments that do not affect the result of later treatments.

Cross-over trials may be recommended when any carry-over effect is short, when the prolongation of the trial neither greatly increases dropout rates nor alters the relative effects of the treatments being compared, when the within-subject variation is less than between-subject variation, and when any order effect can be balanced out.

Warning. It is difficult to prove that the difference between treatment effects is independent of the period of treatment and therefore the Food and Drug Administration in the United States has concluded that the cross-over trial is not the design of choice where unequivocal evidence of treatment effect is required. Hills and Armitage also concluded, "If the number of patients is limited and a cross-over design is chosen, then the internal evidence that the basic assumptions of the cross-over are fulfilled must be presented and if necessary the conclusions should be based on the first period only" [133].

The investigator may, however, be certain from previous studies that the difference between treatments is independent of period and he can then proceed with a cross-over design in certain circumstances. It must be remembered that a difference between treatment effects may be due to a carry-over effect of one treatment into the next period or to an influence of the time of assessment from the beginning of the trial, a so-called order effect.

11.2.1 When any carry-over effect is of short duration

In the treatment of hypertension with antihypertensive drugs the carry-over effect in lowering blood pressure is usually short, and a brief interval between treatments will ensure that one treatment does not influence the result of the next. The interval may last from two-four weeks and has been called a wash-out period. However, certain antihypertensive drugs may have longer effects on measurements other than blood pressure (for example, diuretics may reduce serum potassium for three months). With such long effects the cross-over trial design may not be appropriate, but under certain stringent conditions the trial may still be analysed with the carry-over effect balanced out and even estimated (section 11.5).

11.2.2 When extending the treatment period does not diminish the difference between the treatment effects

As discussed earlier it is possible that the difference between the two treatments may differ at the start of a trial from later periods in the trial. Meier and Free [134] have therefore argued that "each patient as his own control" is not

entitled to the status of dogma. They reviewed the results of cross-over studies on the use of analgesics in postoperative pain. In this situation the pain is lessening with time and the standard design has the advantage of simplicity. However, the between-patient differences are considerable, supporting the use of a cross-over trial, and it is possible to allow or adjust for the order effect in the design or analysis. Although the treatment effects may diminish through time the differences between various treatments may be more consistent.

11.2.3 When a baseline measurement cannot be made

In a cross-over trial baseline measurements may be desirable but they are not essential as all treatment effects are measured within subject. In a parallel-groups trial, precision may be increased when the within-subject variance is lower than that of the between-subject variance and baseline measurements are employed (section 10.8.3). However, if baseline measurements cannot be made, the parallel-group trial will require many more patients than a cross-over trial. One example would be provided by patients with severe diabetes mellitus who require insulin treatment every day. If two new insulins are to be compared for their effect on blood sugar, the baseline blood sugar would be unsatisfactory as the current treatment cannot be stopped and the starting blood sugar represents the efficiency of the original treatment. However, if two diets were to be compared *in addition to* the original treatment then the starting sugar would be a satisfactory baseline measurement.

11.2.4 When a cross-over trial will not result in a large increase in dropouts

The longer an individual takes to complete a trial the more likely that the person will default (dropout). This problem is discussed in section 14.3. The dropout rate will increase with time owing to subjects' moving address, having an intercurrent illness, changing their occupation, and taking holidays. All these possibilities increase with the duration of the trial and, in addition, dropout may occur owing to the subject's becoming intolerant of the number of visits, repeated investigations, or one of the treatments employed. If an adverse effect of a treatment is experienced in the first or second treatment period the subject may be unable or unwilling to take further treatments.

When the treatment period is three months, it will reduce the dropout rate per treatment to give four times a given number of patients one of four treatments for three months than to give the number of patients all four treatments over a one year period. However, it may be difficult to recruit four times as many patients for the standard trial design and the costs of recruitment and initial investigation will be increased.

11.2.5 When an order effect is absent or can be balanced out

The order effect is the change in a measurement according to the period of estimation after allowance has been made for the effect of treatment. In a trial of antihypertensive drugs blood pressure can become progressively lower as

the trial proceeds. The exact mechanisms producing this fall in pressure have not been determined. Initially, pressure falls due to familiarisation with the technique of measurement, the observer, and the surroundings. This effect can be reduced by a prolonged run-in period prior to randomisation. Additional reasons have been suggested for the fall in pressure: an effect of any placebo tablets that are given; a phenomenon whereby an initial lowering of pressure makes blood pressure control easier thereafter; and the removal from the trial of persons whose blood pressure rises leaving a higher proportion of those in whom pressure falls. Whatever the cause of a trend with time the subjects must be randomised to receive the treatments with equal frequency at different times. In figure 11-1 and trial design 2, if equal numbers of patients are randomised to the four sequences, then drug A will be as often given first as second, third, or fourth. Similarly with drugs B, C, and D, the order effect for these treatments is said to be balanced out. The order effect can be estimated by comparing the average results for each period, every interval including an equal number of measurements on the four treatments.

The order effect may be important in trials other than those of antihypertensive drugs or analgesics. In the treatment of diseases with a fluctuating course the trial may be commenced when the condition is at its worst and a subsequent improvement is expected as part of the natural course of events.

11.3 THE STUDY OF TWO OR MORE TREATMENTS SIMULTANEOUSLY: FACTORIAL DESIGNS OR TRIALS TO DETECT AN INTERACTION BETWEEN TREATMENTS

Traditionally, in an investigation the experimenter isolates a number of factors and studies the result of altering one factor while holding the others constant. Fisher considered this doctrine to "be more nearly related to expositions of elementary physical theory than to laboratory practice in any branch of research" [135] and we shall consider the advantages of more complex experiments where two factors (treatments) are given together. The simultaneous examination of more than one treatment allows any interaction between the treatments to be determined. If an interaction is not present the experiment allows an extra estimate of the two treatment effects.

11.3.1 The detection of an interaction between the treatments

An interaction is said to be present when the effect of one factor is different in the presence of another factor. Let us consider a trial where the subjects receive either drug A, drug B, A plus B, or placebo (figure 11-1, number 3). This design is known as a factorial experiment and yields two estimates of the effect of two factors, drugs A and B, one estimate of the drug's effect when given alone and one of its effect when given in combination. Let us suppose that the drugs lower serum uric acid, the mean uric acid after drug A being U_A , after drug B, U_B , after both drugs combined, U_{A+B} , and when on placebo, U_0 . Figure 11-2 provides fictional data for a factorial design. The upper lines

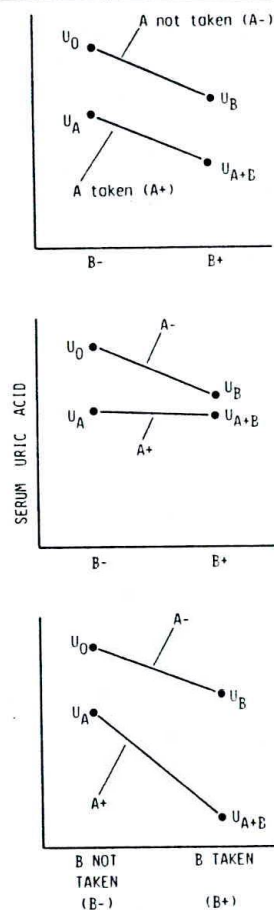


Figure 11-2. The effect of two drugs A and B on serum uric acid. The left-hand results are when B was not taken and the right-hand results when B was taken. The upper lines represent the situation when A was not taken and the lower lines when A was consumed. The upper graph illustrates the situation when no interaction between the treatments was present, the middle when a negative interaction was observed, and the lower graph when a positive interaction was present (see text).

connect the results when treatment A was not taken and the lower lines when it was. The left-hand results were obtained when drug B was not given and the right-hand results when B was taken. The figure illustrates three sets of results: the upper panel when no interaction is present; the middle panel when a negative interaction occurs, and the lower panel when a positive interaction is demonstrated. When there is no interaction the effect of A is the same irrespective of the presence of B and vice versa and the distance between the two lines gives the effect of drug A, equal to $U_A - U_0$ or $U_{A+B} - U_B$. The effect of drug B is $U_B - U_0$ which equals $U_{A+B} - U_A$. When an interaction is not

present the drug effects are said to be additive. In the presence of an interaction the different estimates of the drug effect are not equal.

The middle graph in figure 11-2 illustrates the result when the two drugs in combination have less than the expected additive effect. This is known as a negative interaction and has been loosely referred to as antagonism between the drug effects. The lower graph represents a positive interaction when the effect of the drugs in combination is greater than expected. The term *synergism* has unfortunately been used both for an additive effect (with no interaction) and a multiplicative effect (with a positive interaction). Few trials to detect interactions have been performed; early examples are given by Wilson and his colleagues and Aenishänslin and co-workers [136, 137].

11.3.2 More than one estimate of the treatment effects

The standard design would consist of a group on drug A, a group on drug B, and a group on no treatment. This design would give one estimate of the effect of drug A and one estimate for drug B. Overall, this design yields two estimates for three treatment groups, whereas the factorial design gives two estimates of each drug effect, four estimates for four treatment groups. The factorial design is more efficient and does not lose precision [135]. However, if the effect of drug A is not the same in the presence of drug B (and vice versa) then the factorial design gives only one estimate of the effect of each drug but it does detect the interaction. As discussed previously the estimate of the effect of drug A is given by the two comparisons $U_A - U_0$ and $U_{A+B} - U_B$ and similarly the effect of drug B is given by $U_B - U_0$ and $U_{A+B} - U_A$. The overall estimate of the effect of A is given by

$$\frac{(U_A - U_0) + (U_{A+B} - U_B)}{2}$$

and the effect of B by

$$\frac{(U_B - U_0) + (U_{A+B} - U_A)}{2}$$

This trial design should be more widely used, especially when two treatments are thought to have moderate, but additive, effects (for example, a reduction in mortality of 20 percent for each drug). A trial including both A and B simultaneously may be able to detect a 40 percent reduction in mortality while a standard design of treatment with A alone, B alone, and placebo may fail to detect a reduction of 20 percent.

11.4 CROSS-OVER TRIALS TO DETECT AN INTERACTION

Figure 11-1, design number 4, gives an example of this trial design which is a combination of a within-patient cross-over design and a design to detect an

		ORDER			
		I	II	III	
a)	PATIENT	1	A	C	B
		2	B	A	C
		3	C	B	A

		SITE			
		WRIST	KNEE	METACARPO-PHALANGEAL JT.	
b)	PATIENT	1	A	B	C
		2	B	C	A
		3	C	A	B

Figure 11-3. Two 3×3 Latin squares to allocate each treatment to every patient and to ensure that each treatment is used once at each order or site of administration.

interaction. In such a trial the individual patient has to receive, for example, four different treatments in a certain order. The order can be randomised so that for each set of n patients the order effect and carry effects are cancelled out (section 11.5.3).

Such a design may consist of n treatments arranged in one or two $n \times n$ squares using Latin letters to designate the treatment. Such a design is known as a Latin square and is appropriate for the cross-over design to detect an interaction between treatments and for other trial designs.

The design has been utilized successfully for antihypertensive drugs [101, 138, 139] and antianginal agents [140]. Owing to the lower within-subject variance of measurements of blood pressure and frequency of angina, these trials gave estimates of drug effects with very few patients. However, the trials did not detect interactions between treatments either because they were absent or because the trials were too small to detect these effects.

11.5 THE LATIN SQUARE OR RANDOMISATION SUBJECT TO DOUBLE RESTRICTION (ROW AND COLUMNS)

Figure 11-3 gives two 3×3 Latin squares. The first square gives the order of administration of three treatments A, B, and C to three patients designated 1 to 3, and the second square, for a different trial, gives sites for the application of three treatments. With the first square, for every three patients each treatment is given first, second, and third once only; the design is said to balance out any

order effect. The design also ensures that each patient receives all three treatments. The second square could be useful in a trial of three intraarticular injections in patients with severe generalized rheumatoid arthritis. The most severely affected wrist, knee, and metacarpo-phalangeal joint could be selected, and in the first patient, injection A made into the wrist, B into the knee, and C into the metacarpo-phalangeal joint. For every set of three patients each type of joint will receive all three treatments. The design could be said to be balanced for the joint treated.

11.5.1 Randomisation of patients in a Latin square design

Let us suppose that 18 male patients are to be randomised to a Latin square design with three treatments and three orders of treatment. Three squares identical to a) in figure 11-3 could be taken and three equal to b). These squares can be pooled to an 18×3 table, and the rows numbered one to 18. Eighteen sealed envelopes would be prepared to be opened consecutively as the patients are entered to the trial. In the envelopes would be the row number allocated to a particular patient. For example, the first patient may be randomised to row three, the second to row 17, and so on. The randomisation can be easily read from a random number table (section 7.3) by noting a sequence of numbers less than 19. Randomisation will ensure that the investigator cannot predict the order of treatment for the patients entering the trial. However, he could (but only if he wished) predict the eighteenth order.

11.5.2 The treatment and order effects

The differences between treatments are calculated by comparing the results for the patients when they are taking the particular treatments. Similarly the average result for a particular order (or site) gives the effect of the order or site, provided that each average result for the orders is derived from an equal number of the different treatments.

11.5.3 The carry-over effects

The effects of a treatment may continue into the next period; this is known as a carry-over effect (section 11.1.3). Latin square designs can be employed that balance out residual or carry-out effects [141, 142]. If the number of treatments is even, one square can be designed to achieve this effect, and if the number of treatments is odd two squares are required. Only certain Latin squares have these characteristics. Figure 11-3 gives the two squares required for three treatments. Each treatment follows every other treatment twice when the two squares are employed. Figure 11-4 gives the one square required for four treatments and the two squares necessary with five treatments.

When single squares balanced for carry-over effects are duplicated or a combination of Latin squares is employed that is balanced the residual effects can then be estimated. The methods for calculating the residual and other effects have been clearly described and examples of balanced squares provided for more than five treatments [142].

PATIENT	ORDER			
	I	II	III	IV
1	A	B	D	C
2	B	C	A	D
3	C	D	B	A
4	D	A	C	B

PATIENT	ORDER				
	I	II	III	IV	V
1	A	B	D	E	C
2	B	C	E	A	D
3	C	D	A	B	E
4	D	E	B	C	A
5	E	A	C	D	B

PATIENT	ORDER				
	I	II	III	IV	V
6	A	C	B	E	D
7	B	D	C	A	E
8	C	E	D	B	A
9	D	A	E	C	B
10	E	B	A	D	C

Figure 11-4. Latin square designs to balance out order effects and carry-over (residual) effects. In order to achieve balance, one square is required for four patients and four treatments, and two squares for five patients and five treatments.

11.6 THE GRAECO-LATIN SQUARE

The Graeco-Latin square employs both Latin and Greek letters and allows three different sources of variation to be equalised. For example, a trial may be designed for three drug treatments, three patients, three orders, and three methods of administration (oral, intramuscular, and intravenous). Figure 11-5 gives an example of such a trial. In this Graeco-Latin square each drug treatment A, B, and C is given once orally (α), once intramuscularly (β), and once intravenously (γ). More complex Graeco-Latin squares together with their methods of analysis have been described by Cochran and Cox [142].

11.7 THE SEQUENTIAL TRIAL

Armitage defined a sequential trial as a trial where "Its conduct at any stage depends on the results so far obtained" [143]. Usually a sequential trial compares two treatments, and the results during the course of the trial determine the number of observations made. Frequently subjects are entered to the trial in pairs; one patient is given one treatment and one the other. The results are analysed according to the outcome within these pairs. In most trials of any

PATIENT	ORDER		
	1	2	3
1	A α	B β	C γ
2	B γ	C α	A β
3	C β	A γ	B α

Figure 11-5. A Graeco-Latin square for three patients; three drug treatments A, B, and C; and three orders of administration, α (orally), β (intramuscularly), and γ (intravenously).

design patients are started on treatment serially and not simultaneously; therefore it is a simple matter to assess the response to treatment as it becomes available in sequential order.

A section of statistical theory termed *sequential analysis* derives largely from the work of Wald [144] who allowed for repeated significance testing and derived boundaries describing three possible outcomes. Figure 11-6 gives an example of such boundaries where a comparison is made between drug T and drug A. The upper boundary is a boundary that must be reached to demonstrate a statistical preference for T and the lower boundary must be reached to demonstrate a preference for A. If the two boundaries forming a V shape at the right centre of the figure are reached, then the investigator knows that a preference for one drug over the other is not likely to be demonstrated within the predetermined conditions of the trial. A design with this central limiting boundary is known as a closed design. Figure 11-6 illustrates the result of a trial by Robertson and Armitage [145] where two hypotensive drugs used during operations were compared, T being phenactropinum chloride (Tropenium) and A being trimetaphan (Arfonad). The comparison was made between subjects, and for each pair the time taken for the systolic blood pressure to rise to 100 mm Hg after the use of the drug was measured. The results are plotted and do not reveal a preference for either drug.

Figure 11-7 illustrates the results when two cough suppressants, heroin and pholcodine, were compared with each other and placebo. Comparisons were made within subject [146] and after six days the patients had tried all three treatments and ranked them in order of preference. The trial was designed to

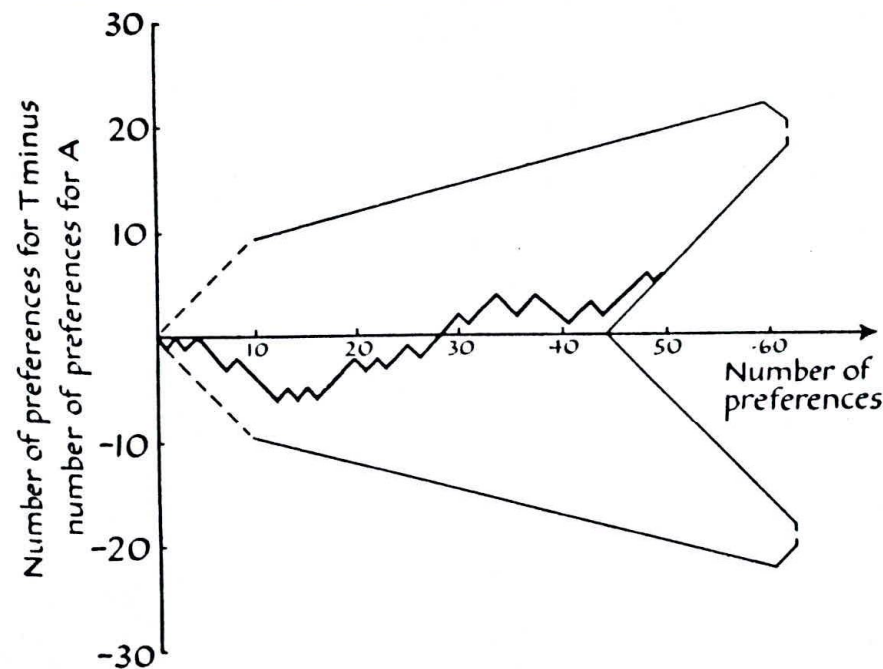


Figure 11-6. The result of a sequential trial to compare two hypotensive drugs employed in anaesthesia, phenactropinum (Trophenium or T), and trimetaphan (Arfonad or A). Reproduced with permission from Robertson and Armitage, *Anaesthesia* 14: 53, 1959.

detect a significant difference between pairs of the treatments at the five percent level of significance with a power of 95 percent when 85 percent of preferences are in favour of one drug. Both heroin and pholcodine were preferred to placebo but no distinction could be made between the two active drugs.

11.7.1 The decisions that have to be made to employ a sequential trial design

As in a standard trial the levels of type I (α) and type II (β) error have to be decided. In addition the trial may be of an open or closed plan. In other words, the investigator has to decide whether he is prepared to allow the sample size to increase indefinitely, or whether he will restrict the trial so that if a specified difference between treatments is not apparent by a certain stage, then the trial is stopped. Figures 11-6 and 11-7 illustrate closed plans. The reader is referred to Armitage [143] for further details. Armitage recommends a closed plan for medical trials as an unexpectedly long series of observations may be a considerable disadvantage. The closed plan reduces the maximum possible sample size.

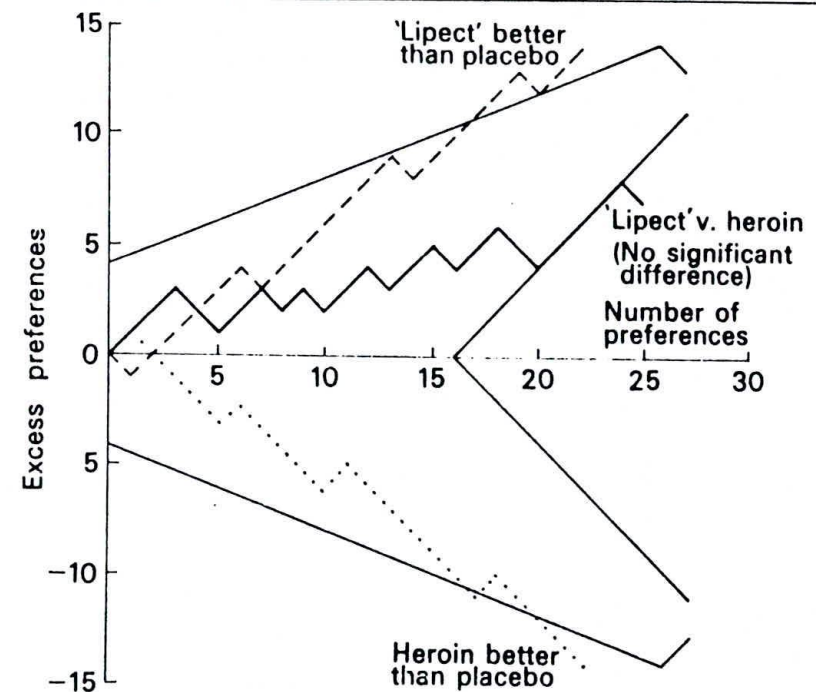


Figure 11-7. A trial to compare pholcodine (Lipect), heroin, and placebo as cough suppressants. Reproduced with permission from Snell and Armitage, *Lancet* 1:860-862, 1957.

- Comparison of Lipect with placebo.
- Comparison of heroin with placebo.
- Comparison of Lipect with heroin.

11.7.2 Advantages of sequential trials

Sequential trials have the ethical advantage of terminating quickly when one drug is an important new advance. These trials may also prove economical and useful as a pilot study to determine the variance of the measurements.

11.7.2.1 Ethical advantages

The investigator may wish to follow the results of the trial closely and continuously and bring the trial to an end immediately when any statistically significant difference is observed: for example, a new treatment for cancer where the treatment is widely available. The investigator will wish to reach a conclusion in the shortest period of time in order to ensure that the treatment, if successful, is generally applied.

11.7.2.2 Economy

If the trial is brought to a speedy conclusion, then the financial cost of the experiment will be reduced. This will be true only when one treatment is much worse than another.

11.7.2.3 Use in pilot studies

Anscombe [147] has suggested that if the numbers required for a nonsequential trial cannot be calculated because the variance of the measurements is unknown, then a pilot study may be conducted in a sequential manner until an estimate of variance has been determined with a given precision. The second stage can then be a standard trial or a further sequential trial.

11.7.3 Disadvantages of the sequential trial

The sequential trial is not suitable for long-term studies or when secondary objectives are important, and the method does not guarantee economy.

11.7.3.1 Difficulty of use in long-term studies

The objective of many sequential trials is to bring the study to an end before many treatments have been started. If the period of the observation is long in comparison to the time taken to enter patients in the trial, then there is little scope for limiting the number of patients who do enter the trial. A sequential trial is most appropriate when the response is obvious soon after treatment is started. Sequential trials are therefore more suitable for the treatment of acute leukaemia than, say, Hodgkin's disease, which has a more prolonged and fluctuating course.

11.7.3.2 Loss of additional information

In the standard trial the larger number of patients may allow end points to be reported with smaller confidence limits and also may allow more observations on related aspects such as side effects or the more severe adverse effects of treatment. A sequential trial is therefore not appropriate when important secondary objectives have been defined.

11.7.3.3 Economy may not be achieved

We must remember that the sequential plan is more economical on average than a nonsequential trial. However, in exceptional circumstances the sequential trial may require more patients and observations than a standard trial. If one treatment is only moderately worse than the other the standard procedure and analysis may be more likely to give a statistically significant result, one final test being performed rather than a repeated sequence of tests (section 10.7.3).

11.7.3.4 Organisational problems

A sequential trial, by definition, will last for an unknown duration. It is therefore more difficult to estimate the total cost of the trial or to know how long to employ staff working on the trial.

11.7.3.5 Concealment of the results during the course of the trial

The statistician in charge of the sequential analysis will naturally plot the results of the trial graphically. If this clear graphical representation is seen by the clinicians involved in the trial, they will naturally have an idea of the likely result of the trial. As a boundary is approached they will imagine that one drug has superiority over another. This conviction will bias their attitude to the trial and may lead to a demand that the trial be stopped. These problems may be overcome by making sure that only a central monitoring committee has access to the results during the course of the trial.

11.7.3.6 Other possible disadvantages

Cochran [148] stated, "In the sequential trial, at the beginning, the doctor is forced to make some decisions about the desired sensitivity of the trial which he can dodge in a fixed-size trial." However, if the researcher is going to estimate the number of subjects he will require for his fixed size trial, then he must make the decisions in the same way as he would for a sequential trial. It is hoped that no investigator will set out on a trial without prior consideration of whether he will recruit sufficient patients for his purpose.

The investigator embarking on a sequential trial will be attracted by the economy in both the subjects involved and the observations required. In addition to the sequential strategy he may also include a low figure for the power term. This will reduce confidence when the boundary is crossed for a nonsignificant difference between the treatments. Although the sequential trial will not differ in this respect from a small nonsequential trial with limited power, there will be a tendency for the trial employing a sequential design to include a lower specification for power. The tendency to use a low-power term in a sequential trial must be avoided. Armitage [143] pointed out that a negative trial that is obviously low in power may inhibit further work, "either because other investigators attach more importance to the first negative results than they deserve . . . or because they have less enthusiasm for repeating previous work than for breaking entirely new ground." Lastly, reporting of the results of a sequential trial should not consist solely of a graphical representation. A summary of the data as a whole must still be presented with means and standard errors, comparisons between relevant subgroups of treatments, and confidence limits.

11.7.4 Conclusions

Despite the difficulties in the design and execution of a sequential trial, the design should be utilized more extensively. If a sequential trial is started and if one treatment proves greatly superior, much may be gained. When the trial fails to reveal a significant difference between the treatments the investigator can calculate the confidence limits of any possible beneficial effect. Armed with

this new knowledge he may or may not proceed to a fixed sample trial of known duration.

11.8 PLAY THE WINNER

The play-the-winner trial has been proposed to limit the number of patients who receive an inferior drug during a clinical trial [149]. A simple example would be to keep using one drug until it first fails and then switch to the second drug until it fails and so on.

11.8.1 Problems with a play-the-winner trial

To my knowledge this method has not yet been tried in a clinical trial and Meier claimed that, "This is testimony to the triumph of good sense over irrelevant theory" [150]. There are three main reasons why the method has not been employed.

11.8.1.1. *The result of the treatment must be unambiguous and known essentially at once*

This is rarely true. In the sequential trial new pairs of patients can be started on treatment prior to the results of the original comparisons being known. This would not be possible in a play-the-winner trial [87].

11.8.1.2 *The population of interest is limited to the group in the particular trial*

Meier considered that if the number of patients receiving the inferior treatment are minimised in the group involved in the trial, it may take longer to get enough of them to come to a reasonably sure conclusion. During this prolonged interval, patients in other centres may receive the inferior treatment and suffer the consequences that the trial seeks to avoid [150].

11.8.1.3 *The investigator may select patients who would respond to a particular drug*

Chalmers [151] suggested that playing the winner will mean that the investigator having successes with drug A will expect the next patient to receive drug A even when he does not know what drug A is. Allocation to treatment is, therefore, no longer blind and Chalmers stated, "It is very easy for a self-fulfilling mechanism to get started in which the winner is ahead and only the winning patients are more and more accepted for the study, thus ensuring that the leader is confirmed as the winner." However, this argument assumes that the investigator can select winning patients for the trial (that is, patients who would respond well to a particular drug). Presumably these patients could have a mild form of disease or other so-called winning characteristic.

11.8.2 Conclusions on play-the-winner trials

A play-the-winner trial is one method of adaptive allocation of patients where the results during the trial determine the treatment to be given to the next patient to enter the trial. A more optimal strategy would be to continuously calculate the probability of success with the two treatments and to allocate

future treatment on this basis—e.g., the so-called two-armed-bandit problem, where the arms of a slot machine are the treatments and inserting a coin is treating a patient [87, 152]. All these procedures assume that all patients are the same; since this is not true, it is safer to randomise and have comparable groups from which conclusions can be drawn. Despite this reservation it would be of interest to see a play-the-winner or related trial performed, and there could be ethical advantages for the investigators involved in this form of trial.

11.9 CONCLUSIONS ON DIFFERENT TRIAL DESIGNS

In this chapter the advantages and disadvantages of the standard parallel groups and cross-over trial designs have been discussed. Cross-over trials tend to be more efficient but the results can be difficult to interpret in the presence of persistent carry-over effects. Factorial designs both within and between subject have been discussed. These designs are very efficient and allow interactions between treatments to be detected. Latin square and Gracco-Latin square designs allow cross-over trials to be performed balanced for order and carry-over effects. Lastly, the advantages and disadvantages of sequential trial designs have been indicated and the concept of play-the-winner trials introduced.

When in doubt, it is safest to employ the standard trial design of one treatment for each person and a fixed number of subjects. However, the numbers required may be greatly reduced by a cross-over or sequential design. When two active treatments are to be tested, a factorial design may well prove the most economical.

12. WRITING THE PROTOCOL

The protocol, or as Bearman preferred, the manual of operations [153], may have to serve many functions: raising monies from a funding agency; obtaining the approval of an ethical committee; recruiting participants; providing a detailed and specific list of instructions on how to perform the trial; and lastly supplying a permanent record of what was intended in the trial. The same document may serve all these functions, and in addition a section on finances may be included when the document is used as a grant application.

The protocol should consist of clear statements on the following: where the trial is being run and by whom; the background of the trial; objectives; numbers to be entered; eligible patients or subjects; procedures to be adopted during the trial; duration over which the trial will be performed; handling of dropouts; proposed analyses; criteria for stopping the trial; publication policy; and financial considerations. A copy of all the documents to be used in the trial should be attached as well.

12.1 WHERE WILL THE TRIAL BE PERFORMED AND BY WHOM?

It may appear self-evident that the personnel involved and the site of the trial should be stated. However, in the case of long-term, often multicentre trials, many difficulties may arise. McFate Smith [154] discussed an organisational model for a multicentre trial and considered that failure to specify the responsibility of various committees has led to difficulties in the execution, analysis,

and publication of large trials. The persons involved in running the trial must be stated and can be grouped as follows.

12.1.1 The steering committee

The steering committee will include the principal investigators and a chairman should be named. This committee should meet with a predetermined frequency and should control the running of the trial.

12.1.2 The coordinating centre

The coordinating centre will include staff with clearly identified responsibilities: clerical staff and staff responsible for quality control, data processing, analysis of results, preparation of reports, and distribution of any medications that may be required. At the centre, persons involved in analysis or publication may be grouped into units or subcommittees.

12.1.3 The clinical centres

The staff involved at the clinical centres must be included in the protocol and acknowledged in any publications.

12.1.4 Advisory board

Most large trials have a panel of experts who constitute an advisory board and whose advice may be sought when necessary.

12.1.5 Funding agency

The source of funds should be stated as should the names of persons responsible for administering the grant.

12.2 BACKGROUND TO THE TRIAL

An exhaustive review of the literature is not called for in the protocol but the reasons for performing the trial should be stated and a few references cited in support.

12.3 OBJECTIVES

The major and minor objectives must be clearly stated and enumerated (chapter 4).

12.4 NUMBER OF PATIENTS REQUIRED

The number of patients or subjects to be entered into the trial in order to achieve the objectives must be stated, together with the assumptions made in arriving at this figure.

12.5 ELIGIBLE PATIENTS

The method of recruiting patients should be noted and the criteria for inclusion and exclusion stated explicitly. It is also desirable to collect a limited amount of

information on those who were considered for the trial but not included and, where possible, details on patients who were available but not considered. The trial participants may then be viewed as a subset of those available and an impression gained of how representative of the population the trial subjects are (section 19.5).

12.6 PROCEDURES TO BE ADOPTED DURING THE TRIAL

This section should include the information to be given to the subjects, the method of obtaining and recording consent, details on any run-in period, how randomisation will be achieved, the treatment schedules to be employed, blindness, data to be collected, and the end points for the trial. Exact methods must be described for any measurements that are to be made (chapter 9). Methods of recruitment, and entry, exclusion, and withdrawal criteria must be stated precisely.

12.7 THE DURATION OF THE TRIAL

An attempt should be made to estimate how long it will take to recruit the required number of patients. The financial support necessary will depend on the expected length of the trial.

12.8 WITHDRAWALS FROM THE TRIAL

The protocol must state under what conditions patients may be withdrawn from the trial and how they should be followed and treated thereafter. The details will include those adverse reactions that require withdrawal and how such reactions, and lesser side effects not leading to withdrawal, are to be detected and treated. It is important to follow all patients who are withdrawn.

12.9 ANALYSIS

The outline proposed for the analysis of the results should be stated and, most importantly, when and how often this analysis will be performed (see section 10.7.3). Details on computer facilities may be required.

12.10 CRITERIA FOR STOPPING THE TRIAL

The criteria under which the trial will be terminated should be stated in as much detail as possible. An attempt should be made to foresee all possible eventualities (section 3.11): for example the mortality from one condition in the intervention group may be reduced with statistical significance, yet total mortality may not be reduced.

12.11 PLANS FOR ORAL PRESENTATIONS AND PUBLICATION

In a large trial involving many investigators a case may be made for planning oral presentations and publications in advance. It may be very destructive if those who are aware of preliminary results leak this information before all the investigators have been informed and before the final report has been agreed

upon. The protocol may therefore state how the final results should be presented to medical colleagues and to the general public.

12.12 FINANCIAL CONSIDERATIONS

Finances should be sought to cover all the costs of the trial. It is not fair to expect a research institution to cover the extra costs of secretarial assistance, computing, stationery, travel, and other overheads. The application for financial assistance should make allowance for these costs and even for monetary inflation if this is expected. However, funding agencies will not usually expect to pay the salaries of the principal investigators or their secretaries; nor will they often be willing to provide office accommodation, typewriters, telephones, and other basic items of equipment that would be expected in a standard office.

12.13 ADDENDA

The following addenda should be included in the trial protocol.

12.13.1 Documents to be used in the trial

Examples of the recording documents should be attached to the protocol.

12.13.2 Details of methods

Full details of all the methods to be used in the trial must be included in the protocol, usually as addenda. Bearman provided an example of a protocol where it was stated that various biochemical measurements must be within normal limits [153]. The protocol did not state the method of performing the biochemical tests nor the normal limits to be expected. These details must be provided.

12.13.3 Quality control

The protocol must state how the precision and accuracy of various measurements will be determined and followed during the trial. If a trial monitor is to be appointed the protocol should state how this person is to perform his or her duties. Details of the training of staff must also be provided.

12.14 SUMMARY OF THE PROTOCOL

For a large trial the protocol will become, of necessity, a bulky and unwieldy document. Each section should be carefully summarised and the protocol be introduced by a brief report of its contents.

12.15 CONCLUSIONS

The protocol must be written with great care. An inadequate document, with insufficient information and sometimes containing errors, is unlikely to attract financial support and may not meet with the approval of an ethical committee. Also, the protocol cannot be employed as an adequate manual of operations if

it lacks the necessary detail. Many large research centres in the United States employ professional writers to finalise protocols, especially when they are used in a grant application.

Appendix 12.16 is based on the headings in chapter 12 and allows the investigator or reviewer to check that the protocol has covered most of the important items in the design of a trial. Similar checklists have been prepared by Spriet and Simon [155] and the Clinical Trial Unit of the London Hospital [156]. Undesirable responses to the questions are in italics.

APPENDIX 12.16

12.16.1 Checklist to assess the protocol for a randomised controlled trial. An undesirable response is in italics.

A1 Who will perform the trial and where?

- | | | |
|---|---------------------------------------|-----------------------------|
| 1.1 Are all the participants named? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 1.2 Are their addresses given? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 1.3 Are steering and other committees necessary? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 1.4 Is the constitution of these committees stated? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| | Not required <input type="checkbox"/> | |

A2 Background to the trial

- | | | |
|---|---------------------------------------|-----------------------------|
| 2.1 Has ethical-committee approval been given? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| | Not required <input type="checkbox"/> | |
| 2.2 Are the authors aware of the literature in their field? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 2.3 Are the authors aware of similar trials that have been completed? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| or that are in progress | Yes <input type="checkbox"/> | No <input type="checkbox"/> |

A3 Objectives

- | | | |
|--|------------------------------|-----------------------------|
| 3.1 Is the major objective clearly stated together with its magnitude? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 3.2 Is the objective realistic? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 3.3 Does the trial answer an important question? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |

A4 Number of subjects required

- | | | |
|--|------------------------------|-----------------------------|
| 4.1 Have the numbers been calculated? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 4.2 Size of type I error <input type="checkbox"/> and type II <input type="checkbox"/> | | |
| 4.3 Is the type I error two-tailed? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 4.4 Can the authors recruit this number of patients? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |

A5 Selection of subjects

- | | | |
|--|------------------------------|-----------------------------|
| 5.1 Is previous treatment allowed? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 5.2 Are the diagnostic criteria clear? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |

Appendix 12.16 (continued)

- | | | |
|--|------------------------------|-----------------------------|
| 5.4 Are selection criteria (age, race, gender) well thought out? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 5.5 Do exclusion criteria cover all ethical problems? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |

A6 Conduct of the trial

- | | | |
|---|---------------------------------------|-----------------------------|
| 6.1 Is the design: parallel groups? <input type="checkbox"/> | | |
| cross-over? <input type="checkbox"/> | | |
| interaction? <input type="checkbox"/> | | |
| sequential? <input type="checkbox"/> | | |
| single-blind? <input type="checkbox"/> | | |
| double-blind? <input type="checkbox"/> | | |
| 6.2 If the trial is cross-over, are there data showing no change in treatment effect with period? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 6.3 Is there a washout period? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 6.4 Will there be a pilot trial? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 6.5 Will blindness be preserved? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| | Not relevant <input type="checkbox"/> | |
| 6.6 Will the observers be adequately trained? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 6.7 Are the end-point measurements valid and repeatable? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 6.8 Are randomisation procedures efficient? | <input type="checkbox"/> | |
| too complicated? | <input type="checkbox"/> | |
| open to manipulation? | <input type="checkbox"/> | |
| 6.9 Is the treatment fixed? <input type="checkbox"/> | | |
| or to be titrated? <input type="checkbox"/> | | |
| 6.10 Are the doses reasonable? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 6.11 Is the labelling and checking of any drug treatment adequate? <input type="checkbox"/> | | |
| or inadequate? <input type="checkbox"/> | | |
| 6.12 Is it clear which accessory treatments will be allowed? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 6.13 Will compliance with treatment be determined? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 6.14 Is the subsequent treatment for patients who complete the trial stated? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 6.15 Will a trial monitor be appointed? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |

A7 Duration of trial

- | | |
|---|---|
| 7.1 Duration for the individual patient | <input type="checkbox"/> weeks/months/years |
| 7.2 Duration of recruitment | <input type="checkbox"/> weeks/months/years |

A8 Withdrawals from the trial

- | | | |
|--|------------------------------|-----------------------------|
| 8.1 Will patients who are withdrawn be followed? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| 8.2 Will severe adverse reactions be detected quickly? | Yes <input type="checkbox"/> | No <input type="checkbox"/> |

Appendix 12.16 (continued)

A9 Analyses

- 9.1 Are the proposed analyses sensible? Yes ☐ No ☐
- 9.2 Will the trial be analysed on the intention-to-treat principle? ☐
the per-protocol principle? ☐
both principles? ☐
- 9.3 Are any necessary computer facilities available? Yes ☐ No ☐

A10 Criteria for stopping the trial

- 10.1 Are possible outcomes clearly stated? Yes ☐ No ☐
- 10.2 Have decision rules for stopping the trial been defined? Yes ☐ No ☐
- 10.3 Will the data be reviewed efficiently as they accumulate? Yes ☐ No ☐

A11 Presentation of the results

- 11.1 Are regular meetings planned? Yes ☐ No ☐
- 11.2 Are the plans for presenting the results acceptable? Yes ☐ No ☐

A12 Financial considerations (n/r = not required)

- 12.1 Are the following staff available when required:
- | | | | |
|---------------|------------------------------|------------------------------|-----------------------------|
| clinicians? | N/r <input type="checkbox"/> | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| nurses? | N/r <input type="checkbox"/> | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| pharmacist? | N/r <input type="checkbox"/> | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| programmer? | N/r <input type="checkbox"/> | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| statistician? | N/r <input type="checkbox"/> | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
| monitor? | N/r <input type="checkbox"/> | Yes <input type="checkbox"/> | No <input type="checkbox"/> |
- 12.2 Are payments to collaborators
- | | |
|----------------|--------------------------|
| adequate? | <input type="checkbox"/> |
| excessive? | <input type="checkbox"/> |
| or inadequate? | <input type="checkbox"/> |
- 12.3 Is the equipment to be ordered
- | | |
|----------------|--------------------------|
| adequate? | <input type="checkbox"/> |
| excessive? | <input type="checkbox"/> |
| or inadequate? | <input type="checkbox"/> |
- 12.4 Are the costs of overheads, travel, and so on
- | | |
|----------------|--------------------------|
| adequate? | <input type="checkbox"/> |
| excessive? | <input type="checkbox"/> |
| or inadequate? | <input type="checkbox"/> |

Appendix 12.16 (continued)

A13 Supporting documents

- 13.1 Are the trial documents clear? ☐
or ambiguous? ☐
efficient for data-processing? ☐
or inefficient? ☐
- 13.2 Is the information collected *too much*? ☐
sufficient? ☐
or *too little*? ☐
- 13.3 Are the methods of measurement well described or not? Yes ☐ No ☐
- 13.4 Are they good methods? Yes ☐ No ☐
- 13.5 Will quality control be
- | | |
|---------------|--------------------------|
| nonexistent? | <input type="checkbox"/> |
| mediocre? | <input type="checkbox"/> |
| or efficient? | <input type="checkbox"/> |

A14 Summary

- 14.1 Does the summary do the trial justice? Yes ☐ No ☐
- 14.2 Is the trial ethical? Yes ☐ No ☐
- 14.3 Will (would) you fund this trial?
- | | |
|----------|--------------------------|
| No | <input type="checkbox"/> |
| Possibly | <input type="checkbox"/> |
| Yes | <input type="checkbox"/> |

13. INFORMATION TO BE COLLECTED DURING A TRIAL

Information must be recorded before a subject enters the trial, throughout the course of the trial, and at the end. Before the start of a clinical trial it must be documented that the patients have the condition under investigation, that there are no contraindications to their entering the trial, and that informed consent has been obtained. During the course of the trial both the benefits and adverse effects of treatment must be recorded to demonstrate that the patients may safely continue in the trial. At the end of the trial the final data must be recorded. These will include full details on defaulters and the attempts made to contact them. This chapter also considers the quantity of data to be collected, the design of documents, the questions to be asked, and the various stages of data preparation.

13.1 THE QUANTITY OF DATA TO BE COLLECTED

Hamilton has cautioned against collecting too much information: "It would be better to resist the temptation to collect every kind of information and spend the time first in thinking more carefully about what would be relevant, and to devise hypotheses to be tested" [157]. Wright and Haybittle have agreed with this assessment and cautioned that the investigator may be unwilling to enter a patient in the trial if there is a lot of paperwork and also that the quality of the paperwork may deteriorate as the quantity increases [158]. On the other hand, data have to be carefully recorded on the outcome of the trial, both on the benefits and disadvantages of treatment; the initial information must be ade-

quate to identify the patients in the trial and show that the different treatment groups were similar with respect to important starting characteristics. If minor objectives have been defined these will also require the collection of extra information.

13.2 FACILITATING DATA ENTRY

Three strategies will reduce the time spent on data recording by the individual investigator. The patients or ancillary staff may complete some of the documents, and some information may possibly be transferred from one computer to another.

13.2.1 The patient can complete certain documents

The patient can be asked to complete a form giving all the identifying and demographic information discussed in section 13.3. This form can be extended to include past medical history, past treatment, family history, occupation, cigarette and alcohol consumption, and other items relevant to the trial.

During the course of the trial the patient can be asked to complete self-administered questionnaires on symptomatic and general well-being (section 16.5). This strategy may save the investigator time and effort and may improve the quality of the data when subjective symptoms have to be assessed.

13.2.2 Ancillary staff may complete certain documents

Clerical, secretarial, or nursing staff may prove more accurate and conscientious than the investigator when transcribing investigation and other results from the medical records to the trial documents. Whenever possible these tasks should be completed uninterrupted by the urgency of patient consultations.

13.2.3 Direct transfer of computer-held information to the computer-held records for the trial

Biochemical results, electrocardiographic tracings, and haematological findings are often held on computer tapes and, in theory, could be transferred directly to the trial data tapes for final analysis. In fact, however, this remains a hope for the future rather than a common occurrence at the present. The computers that serve these different functions usually originate from different manufacturers and the problems of tape conversion may be formidable. The details of patient identification may also vary between the different tapes and pose extra problems when extracting information from one computer file and including it in another.

13.3 INFORMATION IDENTIFYING THE PATIENT

The record forms used in a trial must be considered as confidential information but in a clinical trial the degree of confidentiality need not be greater than with the usual medical records. It is therefore common practice to include name, address, and other identifying features on the first document to be completed.

Subsequent record forms may include extremely confidential information or be sent through the post with the possibility that they are opened by the wrong person or discarded in a public place. For example, a symptom questionnaire has been mailed that included questions on sexual function in order to detect drug side effects [159, 160]. It is prudent to identify these documents with a trial number alone, and the code should be held by the investigator. The patient should be assured of the confidentiality of the information; this is important as documents have gone astray in hospital postal systems and been discovered in discarded refuse (it is hoped only after the data have been abstracted). The investigator must take great care of all confidential information and the record forms should be shredded before disposal.

If the name and address are only included on the initial trial record form and a code number used thereafter, the use of a single number alone may lead to errors in that the wrong number may be entered on a particular record form. For this reason and when sensitive information is not being recorded, many investigators prefer to have both name and trial number on every document. Similarly, two separate numbers for identification will also limit any difficulty in identification.

13.3.1 Items required for identification

Items required include: (1) full name; (2) address; (3) trial number; and (4) hospital number when appropriate.

13.3.2 Other demographic information required

Other information required includes: (1) sex; (2) date of birth; and (3) race.

13.3.3 Demographic information that may be required if the patient is to be identified in central government records

If a patient in a long-term trial is lost to follow-up it is essential to determine whether or not he has died. In the United Kingdom the Office of Population Censuses and Surveys can inform a bona fide medical research worker, for a small fee, whether a patient is dead or alive and if the person is dead, the office can provide a copy of the death certificate. The tracing of such a patient will be facilitated if the National Health Number of the patient has been documented. Also useful for this purpose is the marital status of the patient and the name of the patient's general practitioner (primary care physician).

13.3.4 Other useful general information

13.3.4.1 Telephone number of patient

The patient's telephone number must be obtained because the patient may have to be contacted quickly.

13.3.4.2 Marital status of the patient

This will enable female patients to be addressed correctly but will only rarely be relevant to the conduct or results of the trial.

13.3.4.3 Name and address of the patient's primary care physician

When the trial is not being conducted by the primary care physician he must be informed of the details of the trial. In the United Kingdom every patient has a general practitioner and his agreement should be sought for the patient to take part in a trial. He must be given the opportunity of objecting to his patient's taking part although his written consent is not usually necessary. In my experience the cooperation of the primary care physician can be of great assistance in a long-term clinical trial. Glaser considered that this cooperation is essential in trials involving employees in a pharmaceutical company as the family doctor may be aware of a condition or circumstance that makes it inadvisable for the subject to take part in an experiment [26].

13.4 DATA TO SHOW THAT THE PATIENT IS ELIGIBLE FOR THE TRIAL

Prior to a patient's entry into the trial it is important to document that he satisfies all the entry criteria and has no contraindication to participation in the trial. These criteria are discussed in detail in section 3.9. It is essential that the trial records confirm that all exclusion criteria were in fact negative and all inclusion criteria positive. Quality control cannot be carried out effectively without this documentation.

13.5 DOCUMENTATION OF INFORMED CONSENT

Normally the signed consent form will be included with the trial records; often a copy is given to the patient. If written consent was not required then a note must be made in the trial documents that verbal consent was obtained on a certain date. If a third party was present at these discussions, this fact should be noted.

13.6 DATA TO BE RECORDED DURING THE TRIAL

As discussed in section 13.1 it is very difficult to decide how much information should be recorded during the course of the trial. Data relevant to the prime objective of the trial have to be fully documented. Subsidiary objectives are usually identified (section 4.3) and data relevant to these must also be recorded. It is also essential to record adverse drug effects and symptom side effects.

13.7 DATA TO BE RECORDED AT THE COMPLETION OF THE TRIAL

At the end of the trial the most important end points are determined: for example, the fasting serum cholesterol in a trial of lipid-lowering treatment. If a patient does not appear for this final visit or any other trial consultation, he must be contacted quickly and arrangements made for him to be seen as soon as possible. In a trial of drug treatment it may be possible to give the patients an extra supply of tablets so that if they miss one visit they can continue with treatment until seen. It is of the greatest importance to trace defaulters as they may have died or been withdrawn owing to some adverse effect of treatment or, if in a control group, due to an adverse effect of inadequate treatment. Such

patients must be contacted and asked to return to see the investigator in order that the reasons for default can be determined. The trial documents must include all the details on these patients including the methods used to recall them, the reasons for default and in the event of death, the date and causes of death.

13.8 THE DESIGN OF THE DOCUMENTS

The investigator must be able to complete any forms quickly and easily. The questions must be clear and unambiguous so that the answers recorded are correct, and the documents should be designed so that the data can be processed accurately and efficiently. The investigator should consider the layout of the forms, the provision of instructions to those who complete them, and the duplication of the materials for reasons of security.

13.8.1 Layout of the forms

All documents should be headed by the name of the institute in which the trial is being performed. If there is any possibility of the form's going astray, the name and address of the investigator should also appear. When the form is to be completed by the patient it is important that a brief note accompany the questionnaire or be printed on it stating who is asking the questions, why they are being asked, and assuring that the answers will be treated with complete confidentiality. The patient must also be instructed how to complete the form (see figure 13-2).

The first items on the documents must be those collected first, as the forms should be completed in the order in which the data are obtained. For example, in a trial of an antidiabetic drug, details of previous treatment will be available first and entered first, results of clinical examination will be available next, and the results of investigations, such as blood tests, will be available later and entered on the end of the form.

When one investigator is completing his own trial documents, presumably he will not be concerned with strategies that increase the proportion of questionnaires that are completed. When patients or many different investigators have to complete the documents, it is important to improve the layout in order to maximise response. The reader is referred to standard texts on this subject [161, 162].

The investigator may not be able to compete with the postal sales technique to increase response, for example, where the recipient is told that he has won a prize. To receive this gift he has only to stick a yes stamp into an exactly matched space and return the document in a postage-paid envelope. (In order to avoid purchasing another item, he may also have to take a psychologically less attractive action, for example, refusing an additional and generous, but not free, offer from his benefactor.) Sales techniques have not yet been widely employed in randomised controlled trials, but if the investigator wishes a questionnaire to be completed, he should consider the aesthetic layout of the

In the last 3 months have you suffered from attacks of the following:

sweating	<input type="checkbox"/> No	<input type="checkbox"/> Yes
palpitation	<input type="checkbox"/> No	<input type="checkbox"/> Yes
confusion	<input type="checkbox"/> No	<input type="checkbox"/> Yes
severe apprehension	<input type="checkbox"/> No	<input type="checkbox"/> Yes
hunger	<input type="checkbox"/> No	<input type="checkbox"/> Yes
sleepiness	<input type="checkbox"/> No	<input type="checkbox"/> Yes

Figure 13-1. Section of a questionnaire given to diabetic patients [166]. The condensed format led to some patients' indicating only positive information and omitting negative answers.

document, provide postage-paid envelopes, and consider whether the questions should be answered by ticking yes or no boxes rather than deleting the incorrect answer or entering a correct answer in freehand.

To avoid errors, a box to be ticked should be sited very close to the appropriate answer.

A more common source of error is to shorten the questionnaire by amalgamating questions into blocks as in figure 13-1. In this example the phrase "In the last 3 months have you suffered from *attacks* of the following" does not have to be repeated but patients may not answer all the questions in such a sequence, often only indicating the positive responses. A longer section asking each question separately would result in a higher completion rate for the individual questions and avoid having to make the assumption that no answer equals not present.

13.8.2 Instructions to person completing the form

The instructions for completing a form should appear on the document to be completed and not on a separate sheet as the latter may not be consulted. When appropriate, the instructions should be adjacent to the question being asked

and should state how the responses are to be recorded. For example, if the correct answer has to be ticked, a cross may cause confusion [158].

13.8.3 Duplication of forms prior to use

A small trial may only involve typing the forms and photocopying the top copies. However, in a larger trial many documents may be required and printing will have to be considered as this has certain advantages over typing. Wright and Haybittle [158] considered that printed words are easier to read than typescript as typewriters do not allocate characters a space commensurate with their size but give each letter equal space. These authors also advised a print size comparable to newspaper text and they warned against the use of capitals alone rather than both lowercase and capital letters. The use of capitals on their own can increase reading time by 12 percent [163].

A stencil, photocopying, or printing will provide the forms to be completed but there remains the problem of duplicating the documents after use. No-carbon-required paper may be useful.

13.8.4 The use of no-carbon-required paper

The documents in a clinical trial are of great importance and the use of no-carbon-required paper will allow all recordings to be made in several copies. The copies should be filed separately as an insurance against damage by fire or water, loss in the post, or theft. The loss of research data by theft will probably be accidental, but imagine that your only set of trial documents might be stolen along with your car! Great care must be taken of the information. I well remember the fate of some data I had laboriously collected and recorded with a fountain pen containing washable ink. Someone left a tap running over a weekend in a laboratory on the floor above my office and, following the deluge of water through the ceiling, the data were more than difficult to interpret! Collect all data in duplicate or triplicate and store the copies separately.

13.9 THE FORM OF THE QUESTIONS AND ANSWERS

The characteristics of a good question for a patient are discussed in section 16.3. The present section considers the most desirable features for questions in a trial document to be completed by an investigator: namely, lack of ambiguity; use of positive terms; ease of comprehension; and necessity for a single response. There is obviously a considerable degree of overlap between this section and section 16.3.

13.9.1 The question and answer options must not be ambiguous

Wright and Haybittle provided a good example of question and answer options [158]. In response to a question on the size of a tumour mass, an answer option was: "Reduction of tumour diameter by less than 50%." The authors suggest the answer would be better phrased as, "smaller but not as small as

half the original diameter." The original answer is a phrase that is not strictly ambiguous but simply difficult to understand.

13.9.2 The question should use positive terms

Clark has shown that positive terms are easier to understand than negative terms [164]. For example, the question, "Is the right first toe *longer* than the right second toe?" is to be preferred to the question, "Is the right first toe *shorter* than the right second toe?"

13.9.3 The question must be easily understood

The question must not include difficult words and must be understood by the least intelligent observer. Never use a long word when a short one will do. Abbreviations should be avoided if they are not familiar to all investigators and the questionnaire should also be grammatically correct.

13.9.4 The question must only require a single response

If an observer is asked to record poor circulation to an extremity, he can be asked, "Is a hand or foot white, blue or cold?" The answer options are yes or no. Later, however, the investigators may regret that they did not record whether it was one hand, both hands, a hand and a foot, or even whether only certain fingers or toes were affected. Similarly they may wish to know whether the affected extremities were white, blue, or cold. A series of questions leading to single responses would be preferable. Alternatively, the investigator can be asked to list the fingers or toes that are white and, similarly, those that are blue and those that are cold. If the question is crucial to the outcome of the trial, the temperature of each extremity should be recorded, as an objective measurement is always to be preferred to a subjective assessment, assuming the objective measurement is both valid and repeatable.

13.10 THE RESPONSE TO THE QUESTION

13.10.1 The observer should be asked to tick a response rather than enter a coded reply

An investigator may be asked to indicate whether the patient is taking an anxiolytic drug or not and, if so, which one. He could be asked to enter the name of the drug, but alternatively he could be asked to tick the name on a complete list of possible drugs or to examine the International Nonproprietary Names for Pharmaceutical Substances Classification [165] and enter the numerical code provided by this classification. The provision of a complete list of drugs will enlarge the document considerably but will have the advantage of reminding the investigator of the names of antianxiety drugs.

There are two problems in entering a numerical code from the international classification. First, the form may have to be completed in a hurry in a busy clinic with no time available for the document to be consulted; second, a

physician-recorder, who rarely performs coding duties, may be less accurate than trained clerical personnel whose main job is to provide this service.

In general, the investigator should be asked to enter the response in freehand for subsequent coding or to tick a list of options rather than enter a coded response.

13.10.2 Enter responses in preprinted boxes

The provision of boxes may or may not improve the completion rate, but they should be used for numerical data in order to facilitate later entry to computer files. It has been shown that entering data directly into boxes may increase writing time by ten percent and reduce legibility by three percent [158]. However, when the data are subsequently converted to a machine-readable form time will be saved and some errors may be avoided. Even when the data cannot be punched directly, as with a ticked response, it is often convenient to include a box to hold this information. The boxes also indicate where an answer is expected.

13.10.3 Order of yes and no answers

When several yes/no answers are required consecutively it is advisable not to have a particular response always stated first as it has been suggested that some respondents prefer to tick the first box irrespective of its contents. Occasionally reversing the position of yes/no answer options may inhibit a repetitive response but may lead to errors when most first answers are, say, yes and suddenly one is no. The no may be ticked in mistake for a yes. The best solution may be to roughly alternate the yes/no and no/yes answer options.

13.11 CODING THE RESPONSES

The information provided on the trial documents has to be analysed; in most trials computer programs will be employed to limit the errors in calculation and allow complex statistical computations to be made. If the arithmetic tasks to be performed are not complex, analysis by computer program may take longer than the use of a calculator as all the qualitative data such as gender have to be transformed into numerical codes that can be read by a machine. However, the investigator usually attempts so many analyses—for example, of many variables in several subgroups—that the use of computer programs is inevitable. The data must then be transferred to computer punch cards, paper tape, or directly entered to a computer file via a visual display unit and an attached keyboard. The data must be coded to a numerical format and identified as requiring transfer to a computer file. The position that the data must occupy in that file must be indicated.

Figure 13-2 gives a section of a questionnaire to be completed by a patient and figure 13-3 gives the instructions for coding the information when the codes are to be written by the investigator in the part of the questionnaire

Please tick (✓) correct box or enter required number. All information will be treated with complete confidentiality.	For use of medical staff only.
Please state Your Full Name	Trial number <div style="display: inline-block; border: 1px solid black; width: 20px; height: 20px; margin-right: 5px;"></div> <div style="display: inline-block; border: 1px solid black; width: 20px; height: 20px; margin-right: 5px;"></div> <div style="display: inline-block; border: 1px solid black; width: 20px; height: 20px; margin-right: 5px;"></div> 1-3
Is your father still living? No <input style="width: 20px; height: 20px; border: 1px solid black;" type="checkbox"/> Yes <input style="width: 20px; height: 20px; border: 1px solid black;" type="checkbox"/>	<div style="display: inline-block; border: 1px solid black; width: 20px; height: 20px; margin-right: 5px;"></div> 4
If 'Yes' how old is he now? <div style="display: inline-block; border: 1px solid black; width: 20px; height: 20px; margin-right: 5px;"></div> <div style="display: inline-block; border: 1px solid black; width: 20px; height: 20px; margin-right: 5px;"></div> Years	<div style="display: inline-block; border: 1px solid black; width: 20px; height: 20px; margin-right: 5px;"></div> <div style="display: inline-block; border: 1px solid black; width: 20px; height: 20px; margin-right: 5px;"></div> <div style="display: inline-block; border: 1px solid black; width: 20px; height: 20px; margin-right: 5px;"></div> 5-7
If 'No' how old was he when he died? <div style="display: inline-block; border: 1px solid black; width: 20px; height: 20px; margin-right: 5px;"></div> <div style="display: inline-block; border: 1px solid black; width: 20px; height: 20px; margin-right: 5px;"></div> Years	<div style="display: inline-block; border: 1px solid black; width: 20px; height: 20px; margin-right: 5px;"></div> <div style="display: inline-block; border: 1px solid black; width: 20px; height: 20px; margin-right: 5px;"></div> <div style="display: inline-block; border: 1px solid black; width: 20px; height: 20px; margin-right: 5px;"></div> 8-10

Figure 13-2. An example of part of a questionnaire, designed so that coding and subsequent punching on computer cards is facilitated.

headed "For use of medical staff only." The rest of the heading of figure 13-2 instructs the subjects in how to complete the form and confirms the confidentiality of the document. The patients are asked to enter their full name but only a trial number is entered into the computer file. Entering the full name is time consuming, takes up space in the computer file, and raises questions of confidentiality.

The coding document illustrated in figure 13-3 instructs the coder to enter the trial number with leading zeros on the left, into boxes 1-3. The coder then follows the instructions for boxes 4 to 10. The answer no to the question, "Is your father still living?" is designated as the figure 1; the answer yes as 2; and no answer is to be coded as 9. Similarly, instructions are given for the coding of father's age when alive, and when dead, his age at death. Alphabetical information can be entered into a computer file but is not so easily analysed; therefore the answer no is coded as 1 and not N.

Column (boxes)

1-3

Enter trial number eg.

30

0	3	0
---	---	---

4

No = 1
Yes = 2
No answer = 9

5-7

Age now if living

eg.

0	6	0
---	---	---

No answer

= 999

Not applicable ie. dead

= 888

8-10

Age at death

No answer

= 999

Not applicable ie. alive

= 888

Figure 13-3. Coding document for section of questionnaire illustrated in figure 13-2.

13.12 IDENTIFICATION OF INFORMATION TO BE TRANSFERRED AND THE POSITIONING OF THE DATA IN COMPUTER FILES

The person transferring the data may be instructed to punch or key all the information included in boxes. In figure 13-2 these boxes are further identified by their inclusion under the heading "For use of medical staff only." The positioning of the data in the computer files is indicated by the numbers adjacent to the boxes. In this example punch cards are to be employed with a certain number of columns and a punch card operator is asked to punch the subject (trial) number in columns 1 to 3, the answer to the question, "Is your father still living?" in column 4 and so on. The most commonly used punch cards have 80 columns and can take 80 numbers.

13.13 DIFFERENT METHODS OF DATA CODING

The transformation of information into a numerical form for entry into a computer file is known as coding. The data can be coded on the document (for

Is your father still living?

4	4
No <input type="checkbox"/>	Yes <input type="checkbox"/>
1	2

Is 'Yes' how old is he now?

5-6
<input type="text"/> <input type="text"/>
Years

If 'No' how old was he when he died?

7-8
<input type="text"/> <input type="text"/>
Years

Figure 13-4. A questionnaire with automatic coding of answers (see text).

example, in the margin); coded and written on a separate sheet; or entered directly to the computer file without a separate coding stage.

13.13.1 The information may be transferred to a separate column on the questionnaire

Figure 13-2 illustrates this arrangement: a right-hand column on a patient questionnaire being reserved for the use of medical staff. The coding is carried out on the same piece of paper adjacent to the response, and the number of transcription errors along with the time taken to code data should be reduced.

13.13.2 The information on the documents can be transferred to a separate sheet

The data can be coded and written on a form or sheet that facilitates transfer to the computer system. When cards are to be punched these sheets consist of 80 columns, one line per card. This method often leads to transcription errors and takes the most time. However, the information on several trial documents can be contained in one punching form and may be more easily entered as a punch

operator will spend less time turning over a number of separate trial documents. With this system, and when using a patient questionnaire, the original documents will lack the column headed "For the use of medical staff only." The questions may appear less complicated to the patient and may possibly be completed more frequently.

13.13.3 The information may be entered directly from the questionnaire

Figure 13-4 illustrates how part of the questionnaire in figure 13-2 could be modified for direct computer entry. The number over the top of the box gives the column number to be entered on a card and the punch operator punches a 1 in column 4 if no is ticked and a 2 if yes is indicated; this instruction is given at the right lower angle of the boxes. With the information on age, the punch operator simply punches the age given in columns 5-6 or 7-8. When no answer is given the punch operator would leave the columns blank but could be instructed to enter 9's for missing data. Unfortunately either way there could be confusion between 'not applicable' and 'not answered'. Also, if three boxes are allowed for age then some patients will record

7	6	
---	---	--

 and others

	7	6
--	---	---

 and this may lead to confusion. The punch operator has to be trained to punch this information in the correct right justified manner.

This system removes the coding phase of the data-processing operation but the greatest problem with direct entry is that the keyboard operator, working at speed, may not notice inconsistencies, errors, and scribbled comments. Trial documents must be scrutinized for such problems, and this can be done conveniently during a coding phase. If the coding phase is omitted, the trial documents will still have to be briefly examined.

13.14 CONCLUSIONS

This chapter has outlined the information that should be documented for an individual subject, before, during, and after the performance of the trial. A compromise has to be made between collecting all the information that could conceivably be of interest and limiting the effort involved by documenting only the most essential data. The design of the documents was discussed, including their layout and the form of any questions. The methods of coding trial data were introduced, and strategies to facilitate data entry and subsequent processing were discussed.

14. THE CONDUCT OF THE TRIAL

This chapter considers some of the practical difficulties in performing a trial and the problems that may arise. These may be divided into those that are foreseen and taken account of prior to the start of the trial and those that are unexpected. Murphy's law states, "If anything can go wrong it will." This law has been attributed to captain E. Murphy, a development engineer, who applied it first to an individual technician saying, "If there is any way to do it wrong, he will" [167].

Owing to the probability of unforeseen difficulties it is desirable to have either a separate pilot trial or to commence the trial in a pilot fashion using a provisional protocol. If the pilot trial proves satisfactory and the protocol only requires minor modifications, the trial may continue; otherwise the initial design is abandoned as impracticable. After considering the pilot trial, this chapter discusses the use of a run-in period, the problem of noncompliance by the subject or the investigator, and the difficulties that arise in stopping a trial.

14.1 THE PILOT TRIAL

It is important that the protocol and design for a large or long-term trial is tested either in a separate pilot trial or during a preliminary period of pilot running during which the protocol is open to amendment. The performance of a pilot trial allows the technique of measurement and treatment to be tested, the optimal treatment schedule to be determined, the administration of the trial to be tested, and the rate of recruitment to be assessed. Also, a preliminary

estimate of the treatment effect may be obtained and the wider implications of the trial become more clear.

14.1.1 The examination and therapeutic techniques are tested

During the pilot trial an examination procedure may prove to be unsatisfactory and can be modified. Similarly a therapeutic regimen can be altered if necessary. In section 9.2 the alterations were discussed that proved necessary in order to measure visual acuity during a trial.

14.1.2 The optimum treatment schedule may be identified

The pilot trial may lead to a modification of the dose schedule for a drug by revealing that a large dose produces an unacceptable excess of side effects. Alternatively the pilot trial may demonstrate that an insufficient dose of the drug is being administered. This can be rectified subsequently in the main trial. When two active drugs are being compared in a long-term trial the pilot trial may confirm that they are being used in equipotent doses with respect to their short-term effects. For example, if two drugs are being compared for the prevention of gout over two years, it will be sensible to select doses that reduce the serum uric acid by a similar amount.

14.1.3 The administration of the trial can be tested

The early stages of a trial will involve the entry of patients, randomisation, and the provision of treatment. A coordinating centre may be responsible for the processing of trial documents, randomisation, and the distribution of drugs. All these administrative functions can be tested in the pilot study.

14.1.4 The rate of recruitment can be assessed

Muench's third law states, "In order to be realistic, the number of cases promised in any clinical study must be divided by a factor of at least ten." The law has two important corollaries: "the length of time estimated as necessary to complete a study must be multiplied by a factor of at least ten" and "the sum of money estimated as necessary to complete a study must be multiplied by a factor of at least ten (without inflation)" [96]. There is more than a grain of truth in Muench's third law and the pilot trial may provide an assessment of the recruitment rate that can be achieved in the main trial.

14.1.5 A preliminary estimate of the efficacy of treatment can be made

Although a trial will be designed to detect a certain magnitude of effect, presumably based on prior information, the pilot trial may be useful in confirming that this effect can be achieved under the conditions of the trial. A pilot trial cannot be expected to detect a given reduction in mortality or morbidity as a long period of observation may be required for such an observation. However, if the trial is intended to alter a risk factor by a given amount

with a view to subsequently altering mortality, the change in the factor can be confirmed in the pilot trial.

14.1.6 The pilot trial may enable the investigator to consider the wider implications of the study

Glaser has stated, "A lesser study of an important question is usually of more value than an excellent study of a trivial question" [26]. The reader may not agree with this opinion, but in a pilot trial the investigator may realise that more important questions could be answered by the trial. He may adjust the protocol accordingly to include more subjects or observations than initially envisaged.

14.1.7 The disadvantages of a pilot trial

A pilot trial may delay the start of the full trial, but this may be overcome by starting the major trial with a period of pilot running. This preliminary investigation may not lead to any major change in the trial protocol, but a minor modification may be most important and occasionally the initial trial protocol will be abandoned. However, the results of a pilot trial may be misleading. Muench's first law states, "No full-scale study confirms the lead provided by a pilot study." Other statements have been made as well, including, "Studies can be called pilot studies and so avoid criticism for poor design" and "Pilot studies are a waste of time, money and effort" [96]. It must be emphasized that the results of a pilot study must be treated with great caution; after all, they are only an initial reconnaissance expedition. However, pilot studies do lead to important changes in protocol, are often invaluable, and are always to be recommended. If the protocol is not altered in any major respect the results obtained in the pilot trial may be incorporated in the final analysis, provided an adjustment for repeated looks is made when necessary (section 10.7.3).

14.2 THE RUN-IN PERIOD

The run-in period for a trial is a period of observation during which subjects are considered for entry to the trial. At the end of the period they are randomised to a treatment group if they prove to be eligible. A run-in period is not obligatory and certain advantages and disadvantages are given in table 14-1, and discussed in the sections below.

A run-in period may be appropriate if a preliminary period of observation is required to prove that the patients have a certain condition or to establish other baseline information. The time spent in the run-in period may also be used to: allow any effect of a placebo to take place and thereby reduce this effect after entry to the trial; detect noncompliance with therapeutic advice and exclude such patients; and allow the patients time to consider whether or not they wish to take part in the trial.

Table 14-1. The advantages and disadvantages of having a run-in period prior to randomisation in a trial.

Advantages	Disadvantages
<ol style="list-style-type: none"> 1. Diagnosis can be ascertained with greater certainty. 2. Allows baseline measurements to be made. 3. Placebo response may be reduced after randomisation. 4. Some noncompliant patients may be identified and excluded. 5. Dropout rates are reduced after randomisation. 	<ol style="list-style-type: none"> 1. Patients in the trial are less typical of the general population of patients. 2. Dropout during the run-in period may increase the total number of defaulters. 3. Ethical problems. 4. Expense.

14.2.1 Establishing that the patients have the condition under investigation

If a patient is only examined on a single occasion, it may be difficult to establish a particular diagnosis with confidence. For example, it may be necessary to repeat measurements on subsequent occasions to prove that the patient has, say, a consistently high blood pressure, fasting serum cholesterol, or blood sugar. In addition, it may be necessary to order further investigations to prove that the condition is not secondary to a pathological process that requires immediate attention and that would exclude the patient from the trial.

14.2.2 Establishing baseline measurements

Before commencing treatment in the trial, baseline measurements may have to be made; this period of investigation can constitute a run-in period for the trial. The establishing of a stable baseline is particularly important when regression to the mean is a problem (section 9.1.4). Failure to establish a baseline measurement is especially important when a placebo effect is present.

14.2.3 Estimating the placebo effect

During the run-in period it may be appropriate to give a placebo in order to assess the response to this treatment. If there is a large response to placebo, this may mask the effect of active treatment. For example, in some trials the dose of a particular treatment is not fixed and has to be increased in a stepwise manner according to the response. This titration or increase in treatment may be difficult if a large early placebo effect is occurring. This situation is illustrated in figure 14-1 where active antihypertensive treatment has to be compared with placebo and measurements of blood pressure are presented in a schematic fashion for a patient whose initial casual diastolic pressure is 110 mm Hg, falling to 100 mm Hg after one month on placebo. This fall is due to a combination of regression to the mean, becoming accustomed to the measure-

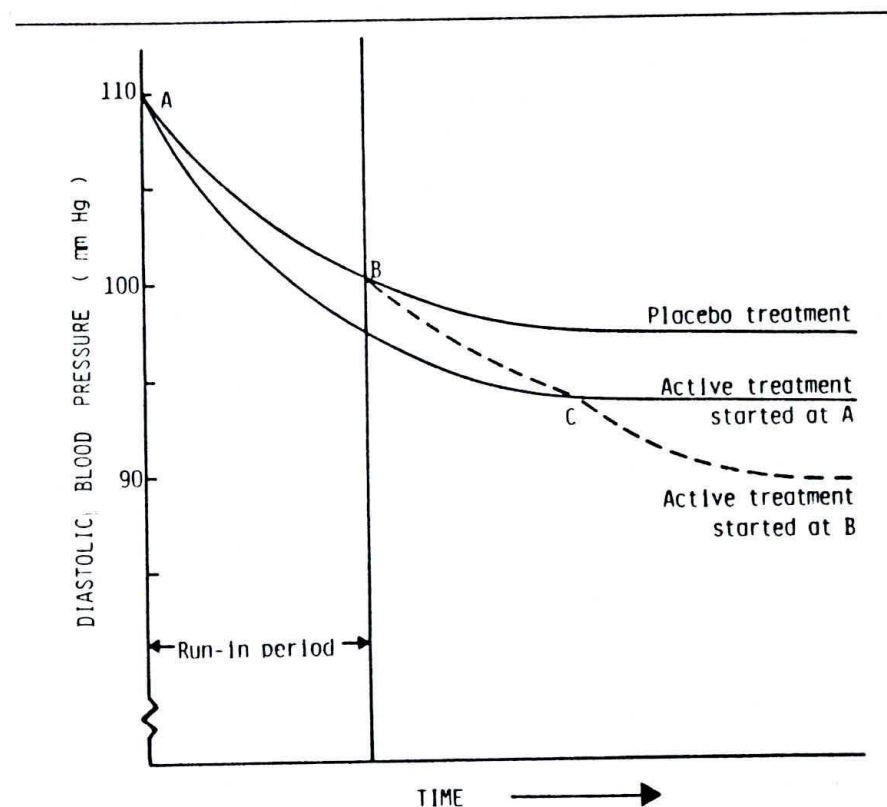


Figure 14-1. A hypothetical trial of the effect of an active treatment and placebo on diastolic blood pressure. When active treatment was started at A, a 15-mm Hg fall in pressure was judged satisfactory at C and the dose of active treatment was not increased. When active treatment was started at B, following a run-in period, a 5-mm Hg fall was not judged satisfactory at C and active treatment was increased (—).

ment, and a possible effect of a placebo tablet. Figure 14-1 also provides the pressure when active treatment is started at A with no run-in period, and at B, after placebo run-in period of one month. When the drug is started at A, the investigator may be content with a 15 mm Hg fall in pressure after one month. On the other hand, when active treatment is started at B the investigator is less likely to be content with a 5 mm Hg fall and may increase the treatment at point C. In this example, the presence of a placebo run-in period results in a wider separation between the effect of placebo and active treatment than when a run-in period is not employed.

Trials with a placebo run-in period and a titration of antihypertensive medication tend to reveal a larger fall in blood pressure on active treatment than trials without a run-in period. The EWPHE trial with a run-in period reported a difference between the active and placebo groups of 25 mm Hg systolic and

12 mm Hg diastolic [168], whereas the MRC trial without a placebo run-in period observed a difference of 15 mm Hg systolic and 7 mm Hg diastolic [106]. There are other differences between the two trials that may account for these results, but the MRC Working Party reported, "Differences in mean pressure between treated and control . . . were less than expected because the fall in pressure in the controls was greater and more prolonged than expected." It is important to start the trial after any nonspecific treatment effect has terminated and the baseline measurement is constant.

14.2.4 Noncompliance can be determined

Noncompliance with therapeutic advice may be determined when a placebo is given during the run-in period (section 14.4). Similarly the willingness of the patients to return for follow-up appointments is tested as well as their compliance with biochemical, radiological, and other investigations. Noncompliant patients may be excluded from entering the trial, but the trial population will be less representative of the whole population as a result.

14.2.5 Dropout may be reduced following randomisation

Default from follow-up during the main course of the trial should be reduced following a placebo run-in period as patients who are unwilling to return for repeated visits or investigations will be excluded. Also some patients become ill soon after starting placebo and attribute the illness to the placebo medication. These patients will usually be excluded or will exclude themselves from the trial. Many of these patients do not experience a coincidental physical illness but are worried about the trial treatment and experience a psychological reaction. Patients who react adversely to placebo medication (and therefore can be expected to react adversely to other treatments) may therefore be excluded. In addition, the run-in period gives the patients time to reconsider whether they are really willing to participate in the trial. Default appears to be more frequent in the initial stages of a trial and a run-in period should exclude a high proportion of trial dropouts.

A run-in period may reduce the degree to which the patients entering the trial are representative of the population as a whole. It may raise the overall default rate, increase the expense of the trial, and pose additional ethical problems.

14.2.6 The trial population may no longer be representative of all patients

Randomisation of subjects eligible for a trial will usually ensure that the different treatment groups in a single trial are similar for important characteristics. However, different trials will include dissimilar patients. Hampton has suggested that similar trials of secondary prevention in myocardial infarction may arrive at different results owing to the unequal characteristics of patients entering the trial [169]. He pointed out that the mortality in the placebo group may provide a clue as to how representative the trial patients are of the population

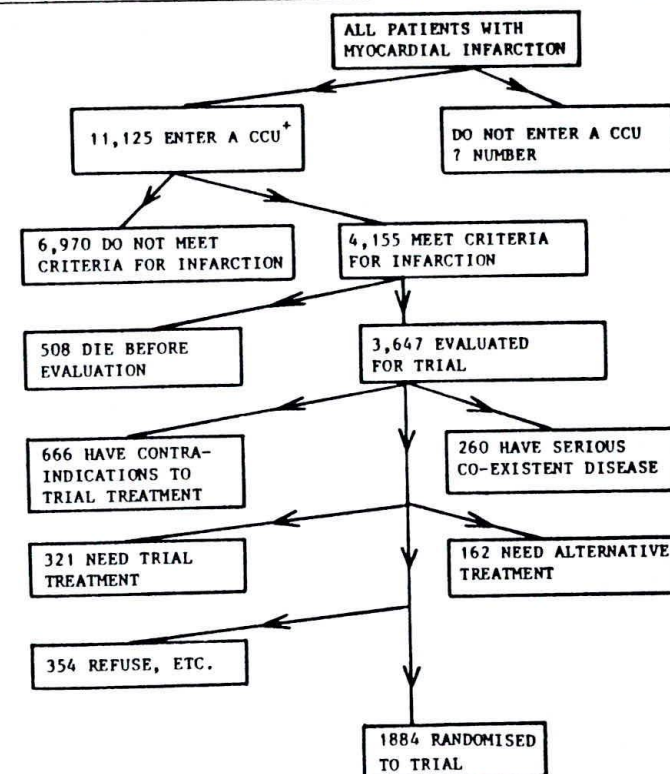


Figure 14-2. Recruitment to the Norwegian Multicentre Trial of Secondary Prevention in myocardial infarction using timolol [170].

*CCU = coronary care unit. Some patients were admitted and evaluated more than once.

of patients as a whole since the usual survival of patients with myocardial infarction is known from prospective population-based epidemiological studies.

It is possible that the results of a trial will depend on the severity of the disease process; it may be difficult to recruit patients who represent the whole community and have an average severity of the condition. Indeed, this may not be desirable if only patients with a given degree of disease severity are considered suitable for the treatment. A recent trial of a beta-adrenoceptor blocking drug in the secondary prevention of myocardial infarction [170] provided documentation both on patients who entered the trial and all patients considered for the trial. This information may be very valuable. Figure 14-2 gives the details for this multicentre trial. An unknown number of Norwegian subjects aged 20-75 years sustained a myocardial infarction, but about 11,000 were admitted to coronary care units during the period of recruitment for the

trial. Of these, less than half met the stringent trial criteria for myocardial infarction and 508 died too quickly to enter the trial. The remaining 3,647 were evaluated for the trial but only 1,884 were randomised. The reasons for not entering the trial were: contraindications to trial treatment; serious coexistent disease; trial treatment necessary; alternative treatment required; and administrative reasons such as the refusal of the patient. At the most optimistic estimate, less than ten percent of the population sustaining a myocardial infarct entered the trial. This shows how difficult it would be to recruit a representative sample.

A run-in period may make the trial subjects even less representative, as the information obtained during this interval may result in the exclusion of patients who do not have all the features of a disease or who are unwilling to attend regularly for supervision and investigation or who do not comply with therapeutic advice. The trial population, unrepresentative of the total population in the first place, is further narrowed to a group who are willing to take advice and medication and to put themselves at considerable inconvenience for the benefit of medical research. This is acceptable as clinical trials are usually intended to be explanatory in nature and examine the effects of treatment in those who receive the intervention. However, it is not surprising that sometimes the results of clinical trials may not apply to the population as a whole.

14.2.7 A long run-in period may result in an increased total number of defaulters

The run-in period may so increase the length of the trial that the default rate from the trial as a whole is increased, even when the dropout rate is reduced for the period of the main trial. However, dropout after randomisation is much more important than default during the run-in period. The latter only reduces the numbers available for the trial but default after randomisation may reduce the comparability of the different treatment groups. If the run-in period reduces the default rate after randomisation, this will be a major advantage (section 4.2.5).

14.2.8 Increased expense

A run-in period will prolong the length of the trial and add to the expense. If a placebo is to be taken during this period, then the cost of this medication and its administration must be taken into account.

14.2.9 Ethical problems with run-in period

It may not be possible to leave patients untreated or to give only a placebo for a period of time. This is discussed in section 3.8. When deciding whether or not to have a run-in period the investigator must examine the advantages and disadvantages for the trial under consideration. When one of the trial treatments is a placebo, a single-blind, run-in period on placebo is usually desirable.

14.3 THE PROBLEM OF WITHDRAWAL FROM A TRIAL

Subjects may be withdrawn from a trial because they default or do not comply with the protocol. They may also die from causes unconnected with the trial or its treatment. The most important patients are those whose withdrawal is related to the trial end points.

At least four percent of patients taking part in a long-term trial may default from follow-up every year despite attempts at recall [37]. The possible number of defaulters can be estimated and the numbers required can be adjusted accordingly. During the course of the trial it is very important to establish that the pattern of default does not differ between the treatment groups. When the reason for default or the proportion of defaulters does differ, omitting them from the analysis may bias the results; the problems of such an analysis are discussed in section 15.7. The results of such a trial must be analysed on an intention-to-treat basis (that is, without omitting defaulters). The alternate analysis when subjects who do not follow the protocol are omitted has been called a per-protocol analysis.

The numbers required for a trial will increase according to the proportion of defaulters. With a trial to be analysed on the per-protocol basis the increase will ensure sufficient numbers to demonstrate a given effect and with an intention-to-treat analysis the proportion of defaulters will indicate the extent to which the effect of treatment may be diluted.

14.3.1 Withdrawals due to reasons unrelated to the end point of the trial

Subjects may be withdrawn when they move address, emigrate, or develop some condition that prevents their further participation in the trial (for example, a fractured spine following a road traffic accident). When calculating the numbers for a trial the number of dropouts must include an allowance for these withdrawals. Such withdrawals should not bias the results of the trial and should be equal in all groups.

14.3.2 Deaths from causes unrelated to the trial end point

The probable number of unrelated deaths can be determined from national statistics. In a long-term trial these numbers should be incorporated in the withdrawal rates, especially in trials of treatment in the elderly. When analysing on the per-protocol basis the rates for unrelated death in the trial must be scrutinised for differences between the treatment groups.

14.3.3 Withdrawal due to criteria related to the trial end points

Withdrawal for reasons related to a trial end point may pose great problems in analysis. The problem can be illustrated by trials of the long-term active treatment of hypertension versus placebo; the end point here is stroke events. We can consider two examples of withdrawals related to the end point: the

development of an adverse effect to active treatment and a deterioration in the hypertensive disease while on placebo.

14.3.3.1 The development of an adverse effect to active treatment

It is possible that patients with either mild or severe disease are more prone to get an adverse effect to active treatment and be withdrawn. In the EWPIIE trial [43] two treatments are employed in the actively treated group. Initially a combination of diuretics is given and if blood pressure control is not satisfactory, methyl dopa is added. Therefore patients with severe hypertension in the actively treated group may be withdrawn owing to adverse reactions to methyl dopa. In the placebo group severe hypertensives are less likely to be withdrawn as a result of placebo methyl dopa treatment. However, there are other reasons why severe hypertensives may be differentially withdrawn from the placebo group (section 14.3.3.2).

It is not always necessary to withdraw the patients because of an adverse effect. For example, hypertensive patients on diuretic treatment may develop diabetes mellitus. The trial protocol can stipulate withdrawal, antidiabetic treatment, or substitution of the diuretic treatment by another active antihypertensive drug.

14.3.3.2 Withdrawal of patients due to a deterioration in the disease process

A trial of antihypertensive treatment in the prevention of stroke may enter untreated hypertensives with a diastolic pressure less than 100 mm Hg. However, there is evidence that patients with diastolic pressures higher than 104 mm Hg require treatment. For ethical reasons it may be decided to withdraw from the trial patients who develop diastolic pressures greater than 104 mm Hg. Such patients will not (we hope) have sustained a stroke but they will not have reached the important trial end point. Moreover, these patients will come almost entirely from the placebo group [98]. As a result of such withdrawals the placebo group may contain progressively fewer severely hypertensive patients as the trial continues. The patients who are withdrawn cannot be said to have reached the stroke end point and yet cannot remain in the trial as it is believed that they may sustain a stroke. Randomisation should have provided groups that were similar at entry to the trial but selective withdrawal will destroy this equality. The problem is so great that one group of workers who failed to observe a stroke in either placebo or actively treated groups wrote [98]:

If withdrawal because of rising blood pressure is regarded as a unsatisfactory outcome then there were many more of these cases in the control than the treated group, but we are doubtful of the validity of this interpretation. Another consequence of these withdrawals is to alter the comparability of the groups. However comparable the two groups may have been at the outset, within 18 months this comparability had disappeared. In the present series the main function of treatment seems to be to prevent the

rise of blood pressure, but it cannot be concluded from this evidence that treatment effectively reduces the risks of death or morbid events due to cardiovascular disease.

14.3.4 Methods for reducing withdrawals

The withdrawal or dropout of subjects from a trial can be reduced by a run-in period to detect patients who react badly to placebos or miss appointments; by excluding patients who have a high probability of withdrawing; by not making the withdrawal criteria too stringent; and by reducing the length of the trial. Withdrawals may also be reduced by maintaining close contact with the subjects during the trial, limiting demands on their time or patience, and continuing efforts at recall when they first miss an appointment.

14.3.4.1 Employ a run-in period

As discussed in section 14.2 the run-in period may be employed to reduce the number of defaulters after randomisation. The dropout rate appears to diminish with time and often subjects agree to enter a trial but do not return for their second visit. These early defaulters will be excluded during the run-in period and if a placebo is given, those who react with many symptoms and cannot take this medication will be excluded.

14.3.4.2 Exclude patients who may have to withdraw

Patients are often excluded from long-term trials if they have a serious coexistent disease that may limit their survival or ability to take part in later phases of the trial. Similarly, subjects who already know that they are going to move to another district, emigrate, or change their occupation are likely to default as the trial progresses; these subjects can be excluded from the trial.

14.3.4.3 The withdrawal criteria must not be too stringent

In a trial lasting for, say, five years, it would not be reasonable to withdraw all patients who missed one assessment or failed to take their medication for a one-week period. Absence of follow-up for six months or no trial treatment for a total of three months would constitute more reasonable criteria for withdrawal. Similarly, it may be reasonable to allow additional treatment to be given during acute illness without withdrawal even if this interferes with the effect of the trial treatment in the short-term. For example, an elderly patient in a long-term trial of antihypertensive medication could be allowed diuretic treatment during an episode of acute bronchitis and not be withdrawn from the trial, even though this therapy will lower blood pressure in the short term.

14.3.4.4 The trial must not be too long

The longer the trial the greater the possibility that the patient will drop out or be withdrawn. The parallel-groups trial design has an advantage in this respect over the necessarily longer cross-over trial.

14.3.4.5 Miscellaneous factors

The subjects must be free to withdraw from the trial at any stage and cannot be asked to commit themselves to taking part for the whole duration. It is likely that the continued enthusiasm and interest of the same investigator may motivate the patient to continue. Factors that reduce default are not well documented, but the patients should not be kept waiting for long periods to see the investigator. They should be given their treatment and investigation without charge and be helped, where necessary, with the cost of transport to the trial centre.

14.3.5 What is the answer to groups made unequal by withdrawals?

Despite the measures discussed in section 14.3.4, withdrawals may occur and the remaining patients may be unrepresentative of the original randomised groups. During the course of the trial the withdrawn patients can be treated in three different ways. First, they can be paired with patients in the other treated group(s) who are also withdrawn; second, the patients withdrawn from one treatment group can be transferred to another; and third, the withdrawn patients can be considered to have reached an important positive or negative trial end point. At the end of the trial, withdrawn patients may be excluded from a per-protocol analysis.

14.3.5.1 Withdrawal of similar patients from the other group(s)

We can consider again the example of active treatment producing diabetes mellitus. Patients developing this condition may be identifiable at entry to the trial (for example, from a high fasting blood sugar). If this is the case, for every patient removed from the active treatment group, a patient could be removed from the placebo group who was matched for initial blood sugar and other characteristics. This approach is full of difficulties as the initial blood sugar may not predict subsequent events and, if it is a good predictor, it would be better to exclude all patients with a high fasting blood sugar from entering the trial.

14.3.5.2 Patients withdrawn from one group should transfer to another group

In a within-patient cross-over trial a patient having to be withdrawn from one treatment can proceed immediately to the next but with loss of data in the first phase of the trial. In a placebo-controlled trial the patients having both a placebo and an unacceptable response would be transferred to active treatment and those in the actively treated group with an adverse effect would transfer to the placebo group. Although this is acceptable for a cross-over trial, with a between-subject design the final treatment groups will become dissimilar with respect to presenting characteristics and the procedure cannot be recommended. However, the authors of the Co-operative Randomised Controlled Trial quoted in section 14.3.3.2 suggested partial treatment for patients with-

drawn from the placebo group owing to severely high blood pressure: "One possible solution might be to maintain the diastolic pressures of the patients in the control group between 110 and 120 mm Hg, or whatever upper limit of diastolic pressure was considered acceptable and compare the deaths and morbid events in this group with those in whom the pressure is maintained below 100 mm Hg." This strategy will compare full treatment in one group with partial treatment in another, hardly a satisfactory solution.

14.3.5.3 Make withdrawal criteria important end points for the trial

In a trial with mortality as an end point it would be inappropriate to equate an adverse effect of treatment with death (for example, the development of diabetes mellitus is unlikely to be fatal). Also when stroke is the end point, not all patients withdrawn because of a high blood pressure would have a stroke if left untreated. However, some will have a stroke and if the incidence is known from previous observational studies for a particular level of blood pressure, a proportion of those withdrawn can be assumed to have had a stroke. Thus for a given number of withdrawals a number of end points will accrue. To my knowledge, the assumption that some withdrawals reach an end point has not been utilised in the analysis of a trial, presumably as any conclusions would rest on unconfirmed assumptions. However, the results of the trial can be examined by assuming that the proportion of withdrawals leading to an end point is zero, one, or various intermediate values.

14.3.6 Conclusion on withdrawals

Patients may be withdrawn from a trial because they were ineligible to enter and should have been excluded initially; they may be withdrawn for ethical reasons or refuse to collaborate further and default from follow-up. The default rate during a large clinical trial is unlikely to be zero even over a short duration. In long-term trials the default rate will be at least four percent per annum [37]; this fact should be taken into account when computing the numbers required for such a trial.

The treatment may precipitate default and defaulters must not be omitted from the first analysis. If analyses are conducted according to the initial randomisation on the intention-to-treat basis, the groups will maintain their initial comparability. However, the interpretation of the results may remain very difficult if withdrawal results in a change of treatment. Also, the investigator may be only interested in the effect of treatment in a subset of the subjects (for example, compliant patients). It is obviously unsatisfactory to assess the effect of dietary advice on, say, lowering blood fats when a proportion of patients refuse to adhere to the diet. Most trials must therefore also be analysed on the per-protocol basis.

The failure of patients to complete a trial may have serious consequences when the investigator intends to complete a series of Latin or Graeco-Latin squares so that order and carry-over effects can be balanced and calculated.

Often a patient will default and remove the possibility of an elegant and simple analysis (section 11.5). The situation may be retrieved, however, by substituting for missing values using multiple regression or other techniques [171].

14.4 THE PROBLEM OF PATIENT NONCOMPLIANCE

Noncompliance is the failure to adhere to therapeutic advice. Some noncompliant patients may be willing to report this at interview or on a self-administered questionnaire. Those who fail to admit to noncompliance cannot be identified by the physician but may be detected by pill count or by more objective measurements.

Failure to comply with therapeutic advice may have important consequences in randomised controlled trials. If the treatment regime is only adhered to by a proportion of subjects in the trial, the results for the patients will relate to the attempt or intention to treat rather than the effect of adhering to the particular intervention. If the patients in the trial are exceptionally compliant the effect of treatment may be large whereas if they are less compliant than expected the effect of treatment may be reduced, too few patients may be recruited or determine the effect, and any dose-response relationship may be underestimated. Lastly we must consider the misinterpretation of trial results that may result from noncompliance and the use of noncompliant patients as a control group.

14.4.1 Interview to detect noncompliance

An interview detects a proportion of noncompliant patients. When noncompliant patients admit to failing to adhere to their treatment, they are almost certainly not lying [172, 173] and an interviewer easily identifies a proportion of such subjects. However, not all noncompliant persons are detected by a simple interview [174, 175, 176], and this method overstates the degree of adherence to a treatment regime. When analysing the results of a trial according to compliance as assessed from an interview we are unable to distinguish between patients who adhere to their treatment and noncompliant patients who refuse to admit to this fact [173].

14.4.2 Self-administered questionnaire to detect noncompliance

The completion of a self-administered questionnaire has the same advantages and disadvantages of the interviewer technique. However, whereas in a small trial there is often ample opportunity for one interviewer to question the patients, if a large number of subjects are involved it is easier and cheaper to arrange for a standard questionnaire on compliance to be completed [177]. A self-administered psychological scale to detect obsessionality has been shown to produce a score inversely related to nonadherence. This may prove useful in detecting patients who are noncompliant [177]. However, such a technique has not yet been attempted in a clinical trial.

14.4.3 Physician's assessment of patients who do not admit to noncompliance

The physician's assessment has been shown to be useless when considering noncompliance with drugs [178]. The physician appears to have no way of predicting who will be noncompliant in the future nor who has been noncompliant during the course of a trial.

14.4.4 Pill count to detect noncompliance

For a pill count the patients are asked to bring their containers of tablets to each visit. The pretext for returning these is to determine whether or not a further prescription is required and the contents are counted out of sight of the patient. The observed number of pills left is compared with the expected number. The method is obviously open to manipulation by the patient and takes no account of the patient's giving other persons his treatment or otherwise destroying some of the tablets [179, 180]. The use of a pill count may overstate the compliance of the patient.

14.4.5 Other measurements of compliance with treatment

Compliance with a drug regime may be tested by measuring drug or metabolite concentrations in the blood or urine. However, this method does not necessarily estimate the day-to-day degree of noncompliance. For example, a patient may usually take his or her tablets but may not have taken them at the time of the blood test. Another patient may omit most of his medication but have taken the treatment just prior to the time the blood sample was taken. Patients may be more likely to take their medication on the day that they make a visit to the investigator, and blood samples at this time may underestimate noncompliance. The number of tablets taken cannot be determined as the pharmacokinetics of many drugs are complex and it is difficult in an individual to predict the concentrations of the drug or its metabolites that would be expected from a given consumption of the drug. If a drug is difficult to detect in the urine, the tablets may be labelled with a fluorescent dye and the urine examined for fluorescence.

Certain drugs have marked effects on blood constituents and these changes may be used to monitor noncompliance. For example, a thiazide diuretic may lower serum potassium and raise serum uric acid. A patient whose serum potassium is not lowered by treatment or in whom the serum uric acid does not change in the expected direction can be suspected of noncompliance [106]. Similarly, in a trial of antismoking advice the carboxyhaemoglobin level in the blood is an indication of whether the patient has stopped smoking or not.

Objective measurements of compliance are to be preferred to indirect methods but may be expensive and relatively more difficult to organise than other measures of compliance.

14.4.6 Owing to noncompliance, later trials of a new treatment may demonstrate a smaller effect than earlier trials

Early trials of new treatments (for example, a new pharmacological agent) will be performed on volunteers or selected patients in a laboratory setting. These subjects are likely to be more compliant than patients subsequently treated outside the research-institute environment. Assuming that noncompliant patients demonstrate a reduced drug effect, the average effect of a given dose of drug will be smaller when these patients are included. Later trials (for example, on outpatients) may therefore suggest a smaller pharmacological effect of the treatment owing to noncompliance.

14.4.7 The numbers required for a randomised control trial may be underestimated owing to noncompliance

When estimating the numbers required for a trial the investigator may employ the results obtained from a pilot or early trial where compliance is high [181]. The numbers required for the trial will be calculated on the basis of the effect shown in this trial; if the effect in the main trial is smaller due to noncompliance, then insufficient numbers may be recruited for the trial.

14.4.8 The dose-response relationship may be underestimated in noncompliant patients

If noncompliance is present, then the effect of a particular drug dose will be underestimated and any estimate of the dose response will be incorrect. In addition, noncompliant patients may appear to take large doses of the drug with no adverse consequences. However, the difference between therapeutic and toxic doses may be small, and when later compliant patients are given the higher doses they may experience an adverse effect.

14.4.9 Noncompliance may lead to a smaller response in one of two equal treatments

In a randomised trial we compare one strategy with another. If one strategy proves more effective than another we may conclude that it is to be preferred, but we cannot necessarily conclude that the pharmacological agent given is more effective. For example, if two drugs are employed in a trial, one may cause side effects and another may not. It is possible that noncompliance will be greater with the drug producing adverse effects than with the other. The drug that induces noncompliance will have its pharmacological effect underestimated and the other drug may appear more pharmacologically active. However, the conclusion from the trial is correct in the sense that the strategy of prescribing the drugs leads to the observed results.

Feinstein provides another example when dietary advice is given [182]. The strategy of dietary advice is compared with no dietary advice and the end point of the trial is survival. If all the cases randomised to diet are compared to all randomised controls, then the results of the trial should be clear and any

benefit from a strategy of suggesting the diet will be apparent. However, it is possible that health-conscious people will comply with the dietary advice and persons who are not worried about their health will not. In this instance, noncompliant patients may continue to smoke and take no exercise and these habits may adversely influence their survival. If the patients who are noncompliant with dietary advice are omitted from the subsequent analysis of survival, then we will compare health-conscious people (who took the dietary advice) with the control patients who consist of both health-conscious and health-careless persons. It would not be surprising if such an analysis proved the diet to be beneficial. A similar example of antismoking advice was discussed in section 8.3. Again, such trials must be analysed on an intention-to-treat basis.

14.4.10 The deliberate use of noncompliant patients as a control group

When a randomised controlled trial cannot be performed it is very difficult to estimate the gain that results from treatment. For example, when considering surgery for a resectable cancer a randomised controlled trial may not be possible but a control group, consisting of persons who refuse to have the operation, may be collected. However, the control group will consist of a very biased and unusual sample; such a study lies outside the scope of this book.

14.4.11 An overall view of noncompliance

Noncompliance by patients during a trial may lead to a misinterpretation of the results. However, the results of a trial that includes noncompliant patients may be more applicable to the effects of the treatment in the community at large. In a randomised controlled trial, procedures should be adopted to detect noncompliance and the results analysed in two ways: first including the noncompliant subjects, and second omitting them. A proportion of noncompliant subjects may be detected by simple and inexpensive interview methods, and urine or blood tests may provide more objective measures of compliance on a single occasion. Noncompliance may also distort any dose-response relationship and lead to inadequate numbers being recruited to a trial.

14.5 NONADHERENCE TO THE PROTOCOL BY INVESTIGATORS

Chapter 12 considered the importance of drawing up a detailed protocol. It is important that all the participants agree to adhere exactly to the protocol. Occasionally the investigator may have to deviate from the protocol in the interest of an individual patient. Such an event will usually lead to the termination of the trial for that particular patient. More important is consistent failure of the investigator to adhere to the protocol. This may not be readily admitted and has to be determined indirectly. Nonadherence can be accidental or intentional and involves three most important areas: admission to the trial of patients who should not be admitted, breaking a double-blind code, and the prescription of additional treatment that is not allowed in the trial protocol.

14.5.1 Admission to the trial of patients who do not fulfil the admission criteria

The coordinating office or the person randomising the patients should confirm whether or not the patients are eligible for the trial. For example, a trial may admit persons with a serum uric acid within a certain range and the coordinating centre must refuse to randomise a patient who does not have a uric acid within this range. An unintentional failure to adhere to the protocol was observed in the University Group Diabetes Program trial where several patients were admitted to the trial without the strict criteria of diabetes mellitus [182]. This problem added to the considerable controversy about this trial, which is discussed in section 19.6.

Patients are occasionally admitted to a trial owing to administrative errors—for example, when a patient is admitted before an investigation result is available. The patient may have fulfilled the other entry criteria to the trial but when the result becomes known it may exclude the patient. Administrative errors may also occur when noncompliant patients are to be excluded by pill counts. A patient may repeatedly forget to return with his tablets and compliance with medication cannot be confirmed.

14.5.2 Breaking a double-blind code

An investigator may deliberately break a double-blind code by examining the details of treatment secured in a sealed envelope. This is unlikely to occur, but he may guess the identity of the treatment from other information. For example, in trials of beta-adrenoceptor blocking drugs the investigator who is prescribing the treatment should not measure the pulse rate as he may detect the marked slowing of the pulse that occurs with this treatment. For this reason one investigator should prescribe and a second investigator assess the results of treatment.

14.5.3 Treatment during the course of a trial

During a trial of drug or dietary treatment other treatments may interfere with the results of the trial and may be prescribed either by accident or design. When comparing active antihypertensive medication with placebo a protocol may not allow the administration of other pressure-lowering drugs. If a patient develops angina, he may require a beta-adrenoceptor blocking drug and have to be withdrawn from the trial as this treatment also lowers blood pressure. The investigator must adhere to the protocol and withdraw the patient.

14.5.4 Conclusions on deviations from protocol

Quality control is critical throughout the trial and should detect deviations from the protocol. Monitors have been appointed in many trials whose task has been to evaluate adherence to the protocol (section 9.2).

14.6 TERMINATING THE TRIAL

In a trial the duration of treatment for the individual patient is specified at the design stage as is the total number of subjects to enter the trial. In the normal course of events the trial will last as long as it takes to enter the required number of patients and for the last patient to complete the study. The trial may be abandoned early if adverse effects of treatment are observed, if a benefit is demonstrated at the predetermined level of significance, or if it is apparent that the required number of patients will never be recruited or that the response to treatment is not as expected.

Decision rules for stopping the trial must be agreed at the design stage and in a long-term trial a review committee must examine the results of the trial at given points in time. The decision rules for stopping the trial will be discussed together with the disadvantages of stopping too early or continuing too long.

14.6.1 Decision rules for stopping the trial

Decision rules are considered in section 3.11 and table 14-2 summarises their general format. The table assumes that interim analyses are performed at predetermined fixed intervals and that the level of significance is adjusted for repeated looks (section 10.7.3).

Rule 1 states that the trial must be terminated if a statistically significant and biologically important adverse effect of treatment is demonstrated. Rule 2 states that a trial must be designed to terminate when any one important

Table 14-2. Decision rules for stopping a long-term trial with morbidity or mortality as an end point.

RULE 1

A statistically significant increase in a serious adverse effect is observed in an actively treated group.

RULE 2

A statistically significant decrease in morbidity or mortality is observed.

PLUS

A reduction in total morbidity and mortality is observed commensurate with there being no transfer of morbidity or mortality from one cause (the end point of the trial) to another.

RULE 3

The predetermined number of patients has been admitted to the trial and followed for a given length of time.

RULE 4

The number of patients recruited will not be adequate or the effect of treatment is not as great as expected.

benefit is demonstrated provided the effect is compatible with an overall benefit to the patients. A reduction in total mortality and morbidity must be apparent, even if not statistically significant as a treatment may reduce, say, myocardial infarction or stroke yet produce an excess of other serious morbid or mortal events. An antihypertensive drug may reduce stroke events but produce episodes of hypotension and an excess of injuries due to falls and other accidents. When terminating such a trial because of a reduction in stroke events, we must be certain that there has been no comparable excess of morbidity from other causes.

It has been suggested that in trials of secondary prevention in myocardial infarction both mortality from myocardial infarction and total mortality must be reduced with statistical significance to stop the trial. However, such a trial may arrive at a statistically significant reduction in myocardial infarction events before total mortality is significantly reduced. It would be unethical to continue the trial until total mortality is also reduced by a commensurate amount.

Rule 3 states that a trial should be terminated when the intended number of patients has been recruited and followed, even when there is a negative result but with the power as specified in the trial design. Rule 4 states that a trial may be aborted if there is no hope of recruiting the required numbers for a given amount of time or money. The trial may also be stopped if the treatment is not producing the effect anticipated in the design. For example, it may be intended to assess the effect on myocardial infarction mortality of lowering serum cholesterol by 1 mmol/l. A diet may be chosen for this purpose put prove to lower serum cholesterol by only half the expected amount. In this event the intended reduction in mortality will be unlikely to be observed and the trial may have to be abandoned.

14.6.2 The disadvantages of stopping the trial too soon

The result of a trial may be accepted if it agrees with the preconceived notions of the medical community and rejected if it agrees with the unexpected (chapter 19). With a surprise outcome it is important to achieve a high level of significance, preferably less than one percent. The pressure to terminate a trial increases as a significant positive result approaches and the temptation must be resisted until the desired level of significance is achieved. One large trial was continued in an attempt to achieve a higher level of significance and will be discussed in section 19.1.

If a trial is intended to prove the efficacy of a treatment that has appeared beneficial in observational studies, then a lower level of significance may suffice. For example, in the EWPHE trial [43] a five percent level of significance has been considered sufficient to stop the trial if stroke incidence is reduced by active antihypertensive treatment. This effect has been shown in middle-aged patients and would not be surprising. However, a novel or unexpected result should reach a higher level of significance.

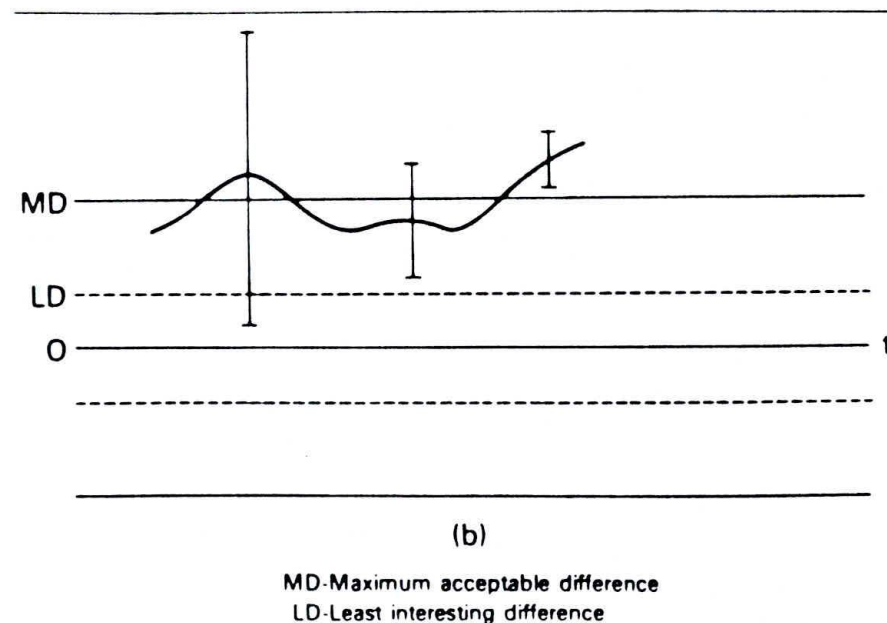


Figure 14-3. The difference between two treatments is plotted against time. The trial will be stopped when the MD is exceeded with statistical significance. Three analyses are performed: at the first, the maximum acceptable difference is exceeded but not with any confidence; at the second analysis, the result lies between the LD and the maximum acceptable difference; and at the third analysis, MD is exceeded with statistical confidence and the trial is terminated. Adapted from Meier, *Clin. Pharmacol. Ther.* 25: 649-650, 1979 with permission.

14.6.3 The disadvantages of stopping the trial too late

When a trial is terminated and a very high level of significance has been achieved, the question arises whether the trial could have been terminated earlier and the benefit of active treatment given to the control group at that time. For example, in a recent trial of secondary prevention of myocardial infarction there were 98 deaths on the active treatment (timolol) and 152 deaths on placebo ($P = 0.0003$) [170]. It is theoretically possible that some of the placebo deaths could have been prevented if the trial had been terminated earlier when $P = 0.01$. However, the review committee may have only examined the data when $P > 0.01$ and later when $P = 0.0003$ and therefore had no opportunity to stop the trial at an earlier stage.

14.6.4 Generalised decision rules for stopping a trial

Meier has formalised some general stopping rules for a clinical trial [150]. He has defined the maximum acceptable difference (MD) as the largest true difference between treatments that a subject in the trial should be expected to accept and yet continue in the trial. He also defined the least interesting difference

(LD) as a true difference, although not terminating the trial, of sufficient magnitude that when exceeded the difference would "be enough to justify a decision in favour of the winning therapy." Figure 14-3 illustrates the possible use of such rules. At the first analysis the MD is exceeded but not with any confidence. At the second analysis the result lies between MD and LD, but at the third analysis the result exceeds MD with statistical confidence and the trial is terminated.

14.7 CONCLUSIONS

This chapter examined some problems that will be encountered during the conduct of a trial. The investigator should first consider performing a pilot trial or at least starting the trial in a pilot fashion. The advantages and disadvantages of a run-in period were discussed together with the problems of reducing dropout, detecting patient noncompliance, and the failure of investigators to adhere to the protocol. Lastly, the decision rules for terminating a trial were considered.

The problems in execution of a trial should be presented in the report and information should be given both on procedures that worked well and those that did not. The reader must understand that errors will occur in the performance of trials and that the results cannot be discounted on the basis of minor deviations from a protocol. Details of such errors should be reported together with the number of patients considered for the trial, entered in the study, and withdrawn.

The decision rules for stopping the trial must be outlined in advance, although not every contingency can be foreseen and catered for.

15. ANALYSIS OF THE TRIAL RESULTS

Several authors have reviewed the statistical analyses appearing in reputable medical journals. The reports reveal an appalling record of error and incompetence [183-188]. This chapter considers these problems but it is beyond the scope of this book to describe in any detail the statistical methods required to analyse the results of randomised controlled trials. The reader is referred to standard texts such as the books by Armitage [189], Snedecor and Cochran [190], and Petrie [191].

Two strategies have been suggested to improve statistical analyses: first, a statement may be made in the protocol of the analyses that are intended. This can be scrutinised by fund-giving agencies and ethical committees [188]. The investigator writing the protocol must have a grasp of elementary statistics or seek the help of a statistician. The second strategy is to have all articles scrutinised by a statistician prior to acceptance for publication. In the past it appears that papers with a detailed description of statistical methods have been referred for a statistical opinion whereas those with little or no mention of these methods have not. The latter articles have proved to be those that actually needed the statistical review [183].

When writing the protocol and requesting funding for a trial, care must be taken to ensure that adequate arrangements are made for prompt data processing and analyses as it may be unethical to proceed with a trial longer than necessary. Peto and his colleagues also considered prolonged trials and stated, "Collect as much data as possible at first presentation, only data which are

strictly necessary thereafter, and analyse the data you do collect very thoroughly" [192]. The initial data determine whether randomisation has been successful in producing equivalent groups and may also be of use in determining prognostic factors.

The present chapter considers errors in analyses; checks on randomisation; analysis of normally distributed data, proportional data, and survival data; confidence limits; and the problems posed by dropouts.

15.1 COMMON ERRORS DURING ANALYSES

The errors frequently encountered in reports of clinical trials include: confusion about the experimental unit in the trial; the failure to use a statistical test; failure to state the statistical test employed when one is used; the inappropriate use of *t* tests; the reporting of standard deviation instead of standard error of the mean and vice versa; the use of one-sided significance tests for α ; confusion over the meaning of *P*; and failure to analyse dropouts in a reasonable manner.

15.1.1 Confusion about the experimental unit

It is not always entire individuals that get randomised in clinical trials; rather, the subject may be a bodily part, for example an eye, limb, or joints. In patients with rheumatoid arthritis it is possible to randomise joints to one of two treatments and five patients may have, say, 16 joints randomised for treatment. The most extreme example of more randomisations than subjects is the single-person trial. In this type of trial episodes of illness are randomised for treatment (for example, recurrent episodes of hay fever or asthma). Such a trial has been conducted in a patient with suspected myasthenia gravis who received in random order, placebo, prostigmine, and D-amphetamine [193]. This trial was of diagnostic value in this patient but would a trial of treatment in one person be likely to have any general applicability? For example, a trial of hay fever treatment in one patient demonstrates that drug A is statistically significantly more effective than drug B. What can we deduce from this trial? We can only conclude that drug A is to be preferred in this one subject. We cannot conclude that, in general, patients should be given drug A and we are mainly concerned with the overall validity of our results (chapter 5).

When more than one patient is involved in a trial but there are more randomisations than patients we may conceivably get the results given in table 15-1. In this hypothetical trial 11 children had a total of 66 warts but one child had 22 warts and the others had seven or less [194]. Taking a wart as an experimental unit, treatment A cured significantly more warts (62 percent) than B (28 percent), $P < 0.01$ using a chi-squared test. However, the data show that this result was due to the excellent result in the child with 22 warts. If the experimental unit is taken to be a patient, then four children responded more to A than B, two more to B than A, and five responded similarly—hardly an impressive result. In clinical trials the experimental unit must be the subject and the results arranged so that each subject is equally important. In

Table 15-1. The results of a trial of the treatment of viral warts in 11 children. The apparently beneficial results of treatment A were almost entirely due to the effect of this therapy in child number 6.

Child	Number of warts randomised to treatment A	Number of warts cured	Number of warts randomised to treatment B	Number of warts cured
1	3	0	1	1
2	1	1	1	1
3	3	3	4	0
4	2	1	1	0
5	2	2	2	2
6	11	10	11	0
7	4	0	3	3
8	1	0	2	0
9	1	0	1	0
10	3	2	3	0
11	3	2	3	2
Total	34	21 (62%)	32	9 (28%)

any statistical test the degrees of freedom should be slightly less and never more than the number of subjects.

15.1.2 A statistical test is not employed

Contemporary reports of clinical trials usually employ a statistical test as most authors are convinced of the necessity of supporting their conclusions with tests of statistical significance. However, if the result of treatment is very remarkable, such as the recovery of the first patients with tuberculous meningitis to be given streptomycin, then neither a controlled trial nor statistical analysis may be required. Unfortunately most treatments are not so effective, the outcome of the untreated condition is more variable, and both controlled trials and statistical evaluation are required.

15.1.3 The statistical test is not identified

Three reviews have reported that about a third of publications state a *P* value without quoting the procedure used to test for statistical significance [183, 188, 195]. It is of great importance to state the method used to arrive at a conclusion in order that the readers may fully evaluate the results and check the statistical tests.

15.1.4 The statistical test is used inappropriately

Glantz [188] has recently demonstrated that at least a third of articles quoting the results of more than one *t* test should have employed a test that allowed for the use of multiple comparisons.

Let us consider the results given in figure 15-1a. Four drugs are compared: A, B, C, and D. Often six *t* tests are reported in order to test for differences

a)

DRUG	A	B	C	D
Mean Results	\bar{x}_A	\bar{x}_B	\bar{x}_C	\bar{x}_D
	t_1		t_2	t_3
	t_4			
		t_5		
	t_6			

b)

TIME (YEARS)		1	2	3	4
Mean results	\bar{x}_0	\bar{x}_1	\bar{x}_2	\bar{x}_3	\bar{x}_4
Mean change	-	$\bar{x}_1 - \bar{x}_0$	$\bar{x}_2 - \bar{x}_0$	$\bar{x}_3 - \bar{x}_0$	$\bar{x}_4 - \bar{x}_0$
Paired t test	-	t_1	t_2	t_3	t_4

Figure 15-1. Results in two trials where multiple t tests should not be employed.

Figure 15-1a. Gives the mean results (\bar{x}) for four treatments A, B, C, and D. t_1 to t_6 indicates the multiple t tests that could be (but should not be) employed.

Figure 15-1b. Provides the mean results for one group of a within-patient trial given a single treatment for four years. \bar{x}_0 is the mean baseline result. t_1 to t_4 are the four paired t tests that could be performed on the change in the variable. Such multiple testing should be avoided (see text).

between the four mean results. By employing the six tests (t_1 to t_6) all results are compared and yet we know that the extreme results must not be selected and contrasted. When only one t test has to be performed in a test for a five percent level of significance, we may be willing to accept a one in 20 chance that we have incorrectly rejected the null hypothesis. If we perform six t -tests we approach a one in four probability of a wrong conclusion. The probability is not as high as $6 \times 5\% = 30\%$, but is given by the formula:

$$\begin{aligned}
 \text{Probability of making a mistake} &= 1 - \text{probability of not making a mistake} \quad (15.1) \\
 &= 1 - 0.95^6 \\
 &= 1 - 0.74 \\
 &= 0.26
 \end{aligned}$$

If there were ten drug groups it would be more obvious that we cannot select the smallest and the largest result and compare the two using a straightforward t test. The solution to the problem is to select a test that takes into account the multiple comparisons performed: for example, a studentized range test [189], or Scheffé's test [190], or, where appropriate, to perform an analysis of variance [189-191].

Figure 15-1b illustrates the situation where one group in a trial is given a particular treatment and the average result of a variable, say, weight, is \bar{x}_0 at baseline and \bar{x}_1 to \bar{x}_4 after 1, 2, 3, and 4 years, respectively. The investigator wishes to see if there has been any statistically significant change in weight after any one of the time intervals and the change is best examined using a paired t test. Unfortunately this gives rise to four tests of significance, and we have the problem of multiple testing and selecting the extremes. Analysis of variance would be an appropriate solution in order to examine the data for a significant change after any one of the four years. The analysis can even be extended to detect a tendency for weight to alter progressively and linearly with time.

Multiple testing will also be a problem if we examine changes in, say, 20 variables, but confine the results to those at four years. Examining many variables does increase the probability of a significant result, but each test is independent of the others and there is only one way of coping with this problem: demand a high level of significance, say $P < 0.01$.

15.1.5 One-tailed tests of significance

In biological work, the test for a type I (α) error must be two-tailed (chapter 10) as the result may either be in the expected or the unexpected direction. The only exception to this rule is the decision-making trial (for example, a trial to determine whether or not a pharmaceutical company should investigate a new drug). In this instance the firm is not initially concerned with whether or not the new treatment is worse than a control treatment, but only whether the company should investigate the new treatment or not. However, in the usual report of an explanatory trial, if the reader detects a P value based on a one-tailed test this should be multiplied by a factor of two to provide a two-tailed assessment.

15.1.6 The meaning of P

When a statistical test is performed on a difference between two groups, P is the probability that the particular difference would be explained by sampling error and observed if the null hypothesis is true and there is actually no difference. P is therefore the confidence with which we reject the null hypothesis but *not* the confidence with which we accept that the difference is exactly zero. For example, if $P = 0.45$, we can only reject the null hypotheses with a probability of 45 percent (that is, we cannot reject it). Forty-five repeat samples out of 100 would be as extreme as this or greater and 45 percent is the probability of

However, P is not the probability of making any kind of error; we can also make an additional error, type II (β) and support the null hypothesis when there actually is a true difference between the treatments (section 10.6). A large P value does not imply that a true difference has been excluded. Freiman and colleagues [187] examined 71 negative randomised controlled trials mainly published in the *Lancet*, *New England Journal of Medicine*, and the *Journal of the American Medical Association*. Half the trials had more than a 74 percent probability of failing to detect a statistically significant 25 percent improvement with therapy. When reporting the results of controlled trials the confidence limits should be reported for any result (section 15.6).

15.1.7 Standard deviation (standard error) of the mean

The calculation of a standard deviation was described in section 10.4. The standard deviation of a series of normally distributed results summarises the spread or variability of the data. As Glantz stated, "When observations are equally likely to be above or below the mean and more likely to be near the mean than far away, about 95% of them will be within 2 standard deviations on either side of the mean" [191]. A standard deviation therefore summarises the data and is an accepted description of the spread or variability of the raw data. The variability of an average result is given by the standard deviation of the mean which is known as the standard error of the mean (SEM) and is calculated by:

$$\text{SEM} = \frac{\text{Standard deviation of the raw data}}{\sqrt{\text{number observations}}} \quad (15.2)$$

The 95 percent confidence limits for the mean (section 15.6) are calculated by taking almost exactly two standard errors on either side of the mean. Glantz concluded, "... the standard deviation, not the standard error of the mean, should be used to summarise data." This is true when describing the information on subjects before entry to the trial but not necessarily for the results of the trial. The outcome of a trial will usually consist of the difference between the mean results of the separate treatment groups. The variability of this difference will depend on the standard error of the means and not simply on the standard deviations of the raw data.

Assume drug A produces a mean result of \bar{x}_A , standard deviation, SD_A and standard error of the mean, SEM_A . Similarly drug B produces mean \bar{x}_B , SD_B , and SEM_B and a comparison of the two means is made by a t test calculated as follows:

$$t = \frac{\bar{x}_A - \bar{x}_B}{\text{standard error difference}} \quad (15.3)$$

$$t \approx \frac{\bar{x}_A - \bar{x}_B}{\sqrt{SEM_A^2 + SEM_B^2}} \quad (15.4)$$

The value of t could be more accurately calculated by pooling the separate variances [189, 190, 191]. A report of a trial should therefore include the appropriate means and standard errors of the means or standard deviations and the numbers of measurements involved so these calculations can be made.

In a graphical representation of the outcome, a mean result is often represented as $\bar{x} \pm SEM$, and the legend to the figure must always state what the bars represent. Traditionally the figure indicates the mean and one standard error of the mean on either side. It should be noted that this figure does not provide the confidence limits for the mean and it is surprising that such a device has achieved widespread acceptance.

15.1.8 The handling of data on dropouts and withdrawals from the trial

The inclusion or exclusion of withdrawals from the trial often leads to errors in analysis and is discussed in section 15.7.

15.2 ANALYSES TO DETERMINE THAT RANDOMISATION HAS PRODUCED EQUIVALENT GROUPS

All the characteristics of the patients at presentation should be compared between the treatment groups in order to demonstrate that randomisation has produced equivalent groups. For each normally distributed characteristic the mean, number of measurements, standard deviation (rather than SEM as the raw data are being described), and range should be presented. For discrete data the proportions must be reported (for example, the percentage male, married, or black). The appropriate statistical tests must be conducted to see if there are differences between the groups at the start of the study.

The following points should be noted:

1. When a large number of characteristics are compared, one may well differ between the groups by chance alone.
2. When large numbers of patients are entered into a trial, some differences between the groups may be statistically significant but they are unlikely to be large or of biological importance.
3. When the trial includes a small number of patients, significant differences are likely to be of biological importance and even nonsignificant differences may be large.
4. If more statistically significant differences between the groups are detected than would be expected by chance, the randomisation process may have failed and a biased selection into the different groups may have occurred.
5. Any random differences between the groups can be adjusted for retrospectively in the analysis [192].

15.3 ANALYSIS OF THE RESULTS: QUANTITATIVE OR CONTINUOUS DATA

The data must be checked for outliers and distribution, transformed if necessary, and subjected to the appropriate statistical test.

15.3.1 Data checking

It is of utmost importance to rectify any data that have been incorrectly recorded and to eliminate any results that constitute errors in measurement. Great care must be taken that outliers are not removed in order to improve the results and support the investigators' preconceived notions.

The data are checked by examining the frequency distribution for impossible or outlying values. In certain situations, consistency checks can be performed, for example, to confirm that all patients who are pregnant or have gynaecological complaints are female.

When computing, statistical packages such as SPSS (Statistical Package for the Social Sciences) [197] can be used to derive a frequency distribution, mean, and standard deviation. Healy and others have described rules for the detection of outliers [198–200].

15.3.2 Tests to confirm that the data are normally distributed

Normally distributed variables are continuous quantitative measurements (for example, blood sugar) that conform to various tests of distribution. The most important test is the degree of skewness that can be tested for statistically [197] and also examined by plotting the frequency distribution of the variable. An example of a skewed distribution is provided by the frequency of plasma urea that has a right-hand tail of high values. Plasma urea has to be transformed to achieve a normal distribution, but the distribution of blood sugar is not markedly skewed.

The normal distribution is also bell-shaped and a test for kurtosis indicates whether the data have too flat (uniform) or too peaked a distribution.

15.3.3 Transformation of the data

The most efficient ways of comparing continuous data are *t* tests and the analysis of variance. These analyses require a normal distribution but data that is positively skewed (to the right), such as plasma urea, can be rendered more normal by a logarithmic transformation. One line within a computer program can produce the desired transformation and the distribution of the transformed variable examined to confirm that the skewness has been reduced.

15.3.4 The data cannot be transformed to a normal distribution

If the data do not conform to a normal distribution, they can be analysed using nonparametric statistical techniques, often with little loss of efficiency [189–191]. However, statistical techniques based on the normal distribution are most efficient and are to be recommended when possible. When the change in a particular variable is to be analysed, the distribution of the original data may not be important if the change in the variable is normally distributed.

15.3.5 The tests to be used with normally distributed data

The *t* test and analysis of variance should be employed where appropriate (section 15.1.4). The Student's *t* test is a well-known statistical test and the analysis of variance (ANOVA) is becoming more familiar and better understood by the average reader of trial results. ANOVA may prove that the result of one or more treatment groups is significantly different from the other groups. The problem will remain of determining which particular group is significantly different from which other groups. The Student's range and other tests can be used to determine exactly which between-group comparisons are not compatible with the null hypothesis (section 15.1.4).

15.4 ANALYSIS OF THE RESULTS: PROPORTIONAL DATA

Proportional data, such as the percentage cured, improved, or dead, require statistical tests such as the chi-squared (χ^2) test. Figure 15–2a gives the calculation of the usual chi-squared test which incorporates a continuity connection. Tables of χ^2 must be consulted to determine whether or not the result exceeds a given level of significance. When examining the change in a proportion over a period of time in the same subjects we have to perform an analysis suitable for paired data. Figure 15–2b gives the calculation for McNemar's test [189], and the result of the test has to be examined in tables of standardised normal deviates. McNemar's test is useful in determining whether any significant change is occurring within one treatment group rather than comparing the data for differences between groups.

15.5 ANALYSIS OF THE RESULTS: SURVIVAL DATA

The survival from a rapidly fatal disease in two groups can be compared using a chi-square test as in section 15.4. However, when the disease is not quickly fatal we will wish to take into account, not only the fact of death, but also the length of time before death. Also, certain patients may be lost to follow-up, and some allowance has to be made for this fact in the calculations. The appropriate analysis is the construction of a life table.

15.5.1 The life table

A life table is best represented as a graph in which the proportion of survivors over time is plotted. The technique for constructing a life table is well described in Armitage's book on medical statistics [189] and in a useful article by Peto and his colleagues [192]. The data required on each patient for such an analysis include the date of randomisation (not the date on which they developed the medical condition under investigation); the date of completion of the study; whether the patient is dead or alive; if dead, the date of death; and if the patient is lost to follow-up, the date the patient was last known to be alive.

a)	TREATMENT A	TREATMENT B	TOTAL
NO CHANGE	a	b	a+b
CURED	c	d	c+d
TOTAL	a+c	b+d	n

$$\chi^2 = \frac{[(ad-bc) - \frac{1}{2}n]^2 n}{(a+b)(c+d)(a+c)(b+d)}$$

b)

AFTER 1 YEAR

T O T A L		CONDITION +	CONDITION -	
	CONDITION +	K	r	
	CONDITION -	S	m	
				T

McNemar's test

$$U = \frac{r-m}{\sqrt{n}}$$

where $n = r+s$

Figure 15-2. The analysis of proportional data. **Figure 15-2a.** Two different groups of patients receive different treatments. The chi-squared statistic assesses the results of the trial and the answer should be compared to tables of χ^2 on 1 degree of freedom.

Figure 15-2b. A group of T patients is examined to see if there is a change in a condition from entry to trial to a one-year assessment. McNemar's test provides the approximate number of standardised normal deviates (U) that can be examined in the appropriate table.

15.5.2 Statistical test of survival

A statistical test is required to compare the whole of the two survival curves illustrated in figure 15-3. Peto and colleagues have come down firmly in favour of the log-rank test [202, 203] and have described how to perform the test [192]. Statistics such as the median and average survival times can be very inaccurate unless nearly everyone dies and the data are extensive. Treatments may also differ in their acute and long-term effects; for example, the life table may show a benefit for one treatment after one year but this may be reversed after, say, three years. The results in figure 15-3 are straightforward in that the results in the two groups are nonsignificant after five years ($0.3 < P < 0.5$). The confidence limits are discussed in section 15.6.2.

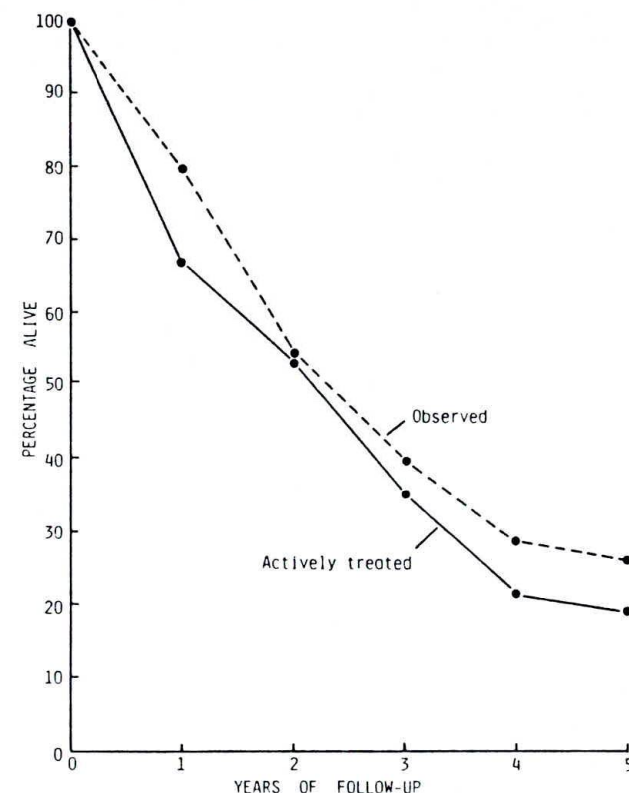


Figure 15-3. Life-table representation for 60 very elderly hypertensive subjects treated with methyldopa and 60 persons who were simply observed [201].

15.5.3 Life-table analysis to determine prognostic features

The life table may also be useful when subdividing the results according to suspected prognostic features but within particular treatment groups. For example, patients given placebo may have two survival curves constructed, one for smokers and one for nonsmokers. The comparison of the curves will indicate the prognostic effect of smoking.

15.6 CONFIDENCE LIMITS MUST BE GIVEN FOR ALL TRIAL RESULTS

Confidence limits are important when a statistically significant result is reported and essential when a negative result is described. If a trial demonstrates a positive result—say, a rise in haemoglobin averaging 1gm/100ml and the result is statistically significant at the 5% level—then we know that within the 95 percent bounds of probability the result was not compatible with a zero increase in haemoglobin. However, the range of increase compatible with a

probability of 0.95 may be large: for example, +0.2 to +1.8 gm/100ml. The limits of this range are known as the 95 percent confidence limits.

15.6.1 The confidence limits for normally distributed data

The confidence limits for normally distributed data are simple to calculate because standard errors are calculated as part of the test of statistical significance. The 95 percent confidence limits for the differences between two treatments in a large number of patients are given by the following formula:

$$95\% \text{ C.L.} = \text{mean difference} \pm 1.96 \times \text{SE (difference)} \quad (15.5)$$

where SE (difference) is the standard error of the difference between the two means. The 95 percent confidence limits are obviously wider than the 90% confidence limits and the latter are calculated as follows:

$$90\% \text{ C.L.} = \text{mean difference} \pm 1.64 \times \text{SE (difference)} \quad (15.6)$$

The importance of reporting confidence limits was stressed by Wulff [204] and supported recently by other authors [122, 205].

15.6.2 Confidence limits for proportional data

For data in a proportional form the standard error of the difference has to be calculated. Baber and Lewis [206] recently reported the confidence limits on a series of trials of the secondary prevention of myocardial infarction. They calculated the SE (difference) and confidence limits as follows:

let n_1 and n_2 be the numbers of subjects in two treatment groups.

let p be the overall death rate and p_1 and p_2 be the death rates in the two treatment groups.

let q be the overall survival rate $= 1 - p$

then

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

$$\text{SE (diff)} = \sqrt{pq \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \quad (15.7)$$

and

$$90\% \text{ C.L.} = (p_1 - p_2) \pm \left[1.64 \times \text{S.E. (diff)} + \frac{1}{2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right] \quad (15.8)$$

where $\frac{1}{2} (1/n_1 + 1/n_2)$ is a correction factor as the data are not continuous.

Confidence limits based on a life table are more difficult to calculate [203,

208, 209]. The 95 percent confidence limits for the negative results after five years reported in figure 15-3 ranged from a benefit due to treatment of 12 percent and an increased mortality of 36 percent.

15.6.3 Confidence limits when the result of the trial is negative

When a trial reports a near-zero effect the confidence limits will indicate the extent to which this result is compatible with a benefit from a treatment in one direction and an adverse effect in the other direction. Of 16 trials with negative results analysed by Baber and Lewis [206] 12 [75 percent] results were compatible with a treatment effect in reducing mortality by 50 percent and eight (50 percent) were compatible with an increase in mortality of 50 percent. Gore has also provided very useful examples of the place of confidence limits in assessing the results of clinical trials, including a study where the confidence limits demonstrated that the trial was needlessly large [207].

15.7 LUMPING AND SPLITTING

Lumping the data together and then analysing them can be contrasted with the practice of splitting the data, presumably into more homogeneous subsections, and then analysing these fractions. McMichael [210] stated, "The aim of a statistical trial is to include all the unpredictable multitude of factors which can influence the outcome by a comprehensive sample. Unless the treatment shows a convincing difference in outcome for the whole group it is not permissible to separate out afterwards a sub-division of better results. Any subdivisions should be done on other criteria before the trial begins." Bradford Hill would agree with the dangers of identifying a subdivision of better results, but when there are factors that influence outcome, he stated "Surely it is our job and duty, to see whether in the analysis we can identify them and thus make them predictable." He added, "It is better to have looked and lost than never looked at all."

The first stage of any analysis must be to examine the groups, as randomised and without any exclusions. This has been termed an analysis on the intention-to-treat principle. The next stage of the analysis may be to allow exclusions and examine groups who followed the protocol, an analysis on the per-protocol basis. Lastly, the data can be analysed according to prognostic and other factors.

15.7.1 Analysis on the intention-to-treat principle

An analysis on the intention-to-treat principle tests the strategy of offering a certain treatment to a group of subjects irrespective of whether they receive or persevere with the treatment. It is the safest way to analyse the results of a clinical trial in order to avoid bias (section 8.6), but it has some disadvantages.

1. In a small trial, default, dropout, or noncompliance may affect only one or two patients but their results may distort the overall result of the trial.

2. In a long-term trial of an active treatment versus placebo treatment, some patients who are intended to receive active treatment may not receive it and some who are intended to receive placebo will receive active treatment. Analysis on the intention-to-treat principle ignores these problems.

In view of these problems analyses should be performed both on the intention-to-treat and on per-protocol basis.

15.7.2 Analysis on the per-protocol basis

An analysis on the per-protocol basis only considers those in the trial groups who received the treatments as specified in the protocol. Patients who do not adhere to the protocol are excluded and not transferred between groups. The worrying effects of such a selection have been discussed in section 8.7, and one trial that has been criticised for concentrating too heavily on this approach is discussed in section 19.1. In this trial it was decided that all events between randomisation and seven days could be ignored as an active treatment only exerts its effect after seven days of treatment [48]. Similarly, events were included for seven days after stopping treatment but not thereafter as the treatment continues to act for only seven days. The results of the trial have not been widely accepted owing largely to the selection of patients and their events for analysis.

15.7.3 Examination of the data according to prognostic and other factors

It is essential for the full analysis of a trial that the data are examined to determine the characteristics of patients in whom the treatment was most effective and least effective. This can only be done for large trials where different subgroups contain considerable numbers of patients. This analysis can be performed even when the divisions were not considered at the design stage and even when splitting the data was suggested by the results. However, the authors should state why they chose certain subdivisions and may be able to report the results with more confidence if they state their intention in the trial protocol. An interesting example was provided by the Hypertension Detection and Follow-up Program (HDFP) trial where patients in the United States of America were randomised to specialist care for their hypertension or referred to the usual community health services for treatment of their hypertension. The overall intention-to-treat analysis showed a benefit from specialist care [132] but when the results were broken down by sex and race, there was no observable benefit for white women [211]. This is an example of a reasonable subdivision of the data yielding important results that may throw further light on the conclusions to be drawn from the trial.

15.8 AMALGAMATING TRIAL RESULTS

The results from different trials may be pooled under the following circumstances.

1. When the condition being treated is similar.
2. When the active or new treatment is similar.
3. When the control treatment is similar (usually a placebo).
4. When randomisation has been employed.

Peto and associates [192] have described how such trials can be pooled by considering each trial as a retrospective stratum of a single large trial. The results for trials of anticoagulants in the secondary prevention of myocardial infarction have been pooled to give a clearer idea of the results of treatment [52].

15.9 CONCLUSIONS

The chapter has considered the common errors to be found in the analysis of trial results, including errors concerning the experimental unit, failure to employ a statistical test, the use of an unspecified or incorrect test, the use of one-tailed tests, and the incorrect interpretation of a negative result. Also outlined were analyses to demonstrate that randomisation has been effective and tests of significance for quantitative and qualitative data. It is essential that the data be checked for errors, the distribution of the data examined, and a transformation performed where necessary. The calculation of confidence limits was discussed and their importance stressed. The analysis of survival data was briefly reviewed and the analysis of data on the intention-to-treat and per-protocol basis was discussed. Other examples of splitting the data and analysing subgroups were presented, and lastly, the amalgamation of results from different trials was mentioned.

16. THE EVALUATION OF SUBJECTIVE WELL-BEING

16.1 INTRODUCTION

When trial end points include reports by the patients on their well-being, the problems of validity and repeatability are greatly increased. Not only do the patients report subjective assessments of their well-being that cannot be confirmed, but different observers may interpret this information in varying ways and record invalid and unrepeatable data.

The recording of a symptom as present will depend on the existence of an organic or psychological cause; the recognition of the symptom by the patient; the reporting of the symptom by the patient; and the recording of the symptom by the observer. These four interrelated stages will be considered separately.

16.1.1 The presence of an organic or psychological cause

The observer, especially if a physician, will usually be concerned about the basis for a symptom. If a symptom does not appear to have an organic cause he may ignore it. However, to the patient a symptom may be equally important whether justified by organic disease or due to psychological problems.

16.1.2 The recognition of the symptom by the patient

Anxious or depressed patients often have a multitude of complaints. Some of these may have an organic basis but would remain unnoticed when the patient

is well, or when recognised, would soon be forgotten. In order for a patient to report a symptom it must be perceived and then remembered.

16.1.3 Reporting of the symptom

When a patient has recalled the presence of a symptom he still has to report the fact to the observer. Both the patient's attitude to the observer and the observer's relationship to the patient may influence whether a symptom is reported or not. For example, if a patient is very grateful to a physician for the care he has received he may be unwilling to report symptoms, viewing such complaints as an expression of ingratitude. On the other hand, a patient with a grievance may possibly list more symptoms.

The attitude of the observer is often crucial in the reporting of symptoms. A sympathetic observer may appear relaxed and encourage conversation in the form of initial small-talk that reassures the patient and leads to a fuller disclosure of complaints. Also the observer may assiduously enquire whether or not the patient has certain symptoms; this may lead to a high rate of reporting. The formidable, brusque observer is unlikely to be told of so many symptoms.

16.1.4 The recording of the symptom by the observer

The observer must report the patients' symptoms without bias. In a clinical trial a physician may take little notice of a symptom that appears to be of a psychological origin, especially if numerous such complaints are reported at the same time. The variation in the frequency of reporting symptom side effects in drug trials is illustrated in table 16-1. The proportion of patients who reported sleepiness in various trials of the antihypertensive but sedative drug, methyldopa, varied between ten percent and 83 percent. Such variability is obviously not acceptable but does occur with current methods of assessment. Table 15-2 also gives details of the control treatment and the percentage complaining of sedation on this treatment. One treatment, clonidine, produced sedation more often than methyldopa but the other control treatments were not known to produce sedation. The difference between the proportion complaining in the methyldopa group and the proportion in the control group provides an estimate of the percentage of complaints attributable to methyldopa. The attributable percentage varied from seven percent to 62 percent and did not appear to be affected by whether the symptom was elicited by direct questioning or arose as a spontaneous report. However, with such variability it is important to standardise the method of collecting these data.

The recording of symptoms in clinical trials is difficult but also of the greatest importance. In a trial to compare two drugs the objective effects may be similar but one drug may produce symptom side effects, thus precluding its use in clinical practice. The remainder of this chapter will discuss standardised methods for eliciting such data. Trials of treatment in psychiatric patients will be reviewed as well. Lastly, the importance of measuring the overall quality of life will be considered and methods for quantifying this concept will be dis-

Table 16-1. The percentage of hypertensive patients complaining of sedation when taking methyl dopa; symptoms assessed by interviewer.

Percentage complaining on methyl dopa = (1)	Control treatment	Percentage complaining on control treatment = (2)	Percentage complaining due to methyl dopa = (1) - (2)	Direct question (DQ) or spontaneous report (SR)	Reference
10	Clonidine	56	?	DQ	212
15	Diuretic/ Placebo	6	9	?	213
24	Bethanidine	12	12	SR	214
25	Debrisoquine	18	7	DQ	215
37	Clonidine	63	?	DQ	216
47	Guanethidine	0.5	46	DQ	217
75	Guanethidine/ Bethanidine	13	62	SR	218
83	Same	51	32	DQ	

cussed. We shall be concerned primarily with recording symptoms and other data of a sensitive and subjective nature. However, the methods and concepts are valid for obtaining any verbal information from a subject.

16.1.5 How to assess the presence or absence of symptoms

The observer may simply record any complaints reported to him, enquire specifically about certain symptoms, or ask the patient to complete a questionnaire without further assistance. The first method can be termed recording of spontaneous complaints; the second, interviewer-administered questioning; and the third, self-administered questioning. The questionnaire methods rely heavily on asking a sensible question. Composition of a good question will be discussed also.

16.2 SPONTANEOUS REPORTING OF SYMPTOMS

When a patient reports a symptom without prompting, this suggests that the complaint is troubling the patient and is a more severe problem than one detected only by direct questioning. It is therefore important to record such information before proceeding to ask specific questions. The difficulty is in standardising the procedure from one patient to the next and from one occasion to another. The observers should be asked to follow the same protocol, for example:

Step one. Ask the patient to sit down, saying, "Good morning (afternoon). Please sit down."

Step two. Ask the patient, "Have you any problems?" This will have to be modified if the patient is known to be suffering from a complaint. The question would then be, "Have you any problems apart from . . . ?"

Step three. The observer must carefully record all the symptoms reported by the patient.

Step four. Code and analyse the information. This may be very difficult owing to the multitude of possible answers and to error or bias in those responsible for coding [219]. In a small trial the responses can be listed and only those analysed that are known to be of interest or reported with any frequency.

Having ascertained spontaneous reports of symptoms, the observer may proceed to ask questions. We must consider what constitutes a good question.

16.3 A GOOD QUESTION

A good question must be easily understood by the subject, limited to an exact period of time, have a high repeatability and good validity when used to detect a clinical condition, have the same meaning for different individuals (be unambiguous), concern only one item at a time, be grammatically correct, and result in an answer. A question may also be open, with the respondent answering in his own words, or closed, where he selects from alternative an-

swers. Any tendency on the part of respondents to always agree either to the first answer or to a positive or negative statement must be discouraged. Certain characteristics of good questions in general were reviewed in 13.9; that section should also be consulted.

16.3.1 The question must be clear

Medical personnel often overestimate their patients' understanding of medical terms and no ambiguous terms should appear in the question. Most trials will include some subjects of lower than average intelligence and the questions should be designed so that these persons can understand and answer. Long words and double negatives should be avoided and whenever possible a difficult long word should be replaced by a short one. Bennett and Ritchie [162] have reviewed the qualities of good questions and considered that leading questions should be avoided where possible and that vague terms such as *occasionally* or *often* must be replaced by precise numerical terms. It is also important to limit the time over which the symptom should be recalled.

Consider the following questions [159].

"Since your last visit have you often felt sleepy during the day?"

Answers: Yes/No

"Have you, in the last three months, noticed weakness in the limbs?"

Answers: Yes/No

The first question uses the term *often* and also leaves the duration rather vague; however, it would be suitable in a trial with a fixed interval between visits. The second question could be used when the visits were three months apart or longer.

16.3.2 The question must have a high repeatability

The measurement of repeatability has been discussed in section 9.1 and indices have been provided that can be applied to the questionnaire answers. To assess repeatability, a group of persons should be asked the same question twice, with the occasions separated by a short interval, say, two weeks. The subjects answering the questions should be from a similar background to those who will be recruited to the trial. The questions must not be repeated after a very short interval, lest the subjects remember their first replies; a number of questions can be tested simultaneously in order to limit this recall. A question that is highly repeatable may be well understood but not necessarily valid: the question may not measure what is intended.

16.3.3 The question must be valid

Validity has been discussed in chapter 5 and can be defined as the extent to which the question measures what it is supposed to measure. An example can be provided by a series of questions that were designed to detect the symptoms

that occur with a marked fall in blood pressure. Certain antihypertensive drugs produce these symptoms when the subject stands up. Three of the questions were as follows [159]:

"Since your last visit have you suffered from unsteadiness, light-headedness or faintness?"

Answers: Yes/No

"Does the unsteadiness or faintness occur only when you are standing?"

Answers: Yes/No

"For how many hours in the day are you troubled by unsteadiness or faintness?"

Answers: Less than one hour/one–two hours/more than two hours

When the blood pressure of the respondents was measured standing and lying, a larger than average fall in pressure on standing was observed when positive answers to the first two questions were reported together with a duration of less than one hour.

This series of questions had a degree of validity but one question that did not appear valid on a preliminary analysis was included in a series of questions to detect a previous history of stroke. The question was

"Have you ever had, without warning, sudden loss of power in an arm?"

Answers: Yes/No

Most of the patients who responded positively to this question had not had a cerebrovascular accident at any time. The false positive rate was therefore high and this question was not acceptably valid. Perhaps the question, "Have you ever had a stroke?" would be more suitable. However, this direct question may still have a high false positive rate owing to confusion with other episodes of illness and a high false negative rate if minor cerebrovascular accidents are to be detected. Medical terms are frequently not understood [220] but may be better understood by those who have the condition that is to be detected [221]. In general, medical terms should be avoided.

16.3.4 The question must not be ambiguous

A question was designed to detect the presence of diarrhoea and was phrased as follows [159]:

"Are your motions often loose or liquid?"

Answers: Yes/No

Faeces are often referred to as motions in colloquial English, but this attempt to use simple words may have led to ambiguity. The question could be interpreted as an enquiry about physical mobility and should be modified to prevent this possible error.

16.3.5 The question should only make one enquiry

The first question in section 16.3.3 asks whether the patients have had unsteadiness, light-headedness or faintness. Patients answered yes if they had suffered from any one of these symptoms or indeed all three. It would be preferable to ask about each symptom separately as it is possible that faintness results from postural hypotension (the condition that is to be detected) whereas unsteadiness may be due to arthritis of the legs or vertigo. If the latter is true, the inclusion of the symptom unsteadiness in the question may increase the false positive rate.

16.3.6 The question should be grammatically correct

It is more important that the question is well understood by the man in the street than whether or not it is grammatically correct. A standardised, repeatable, and widely used questionnaire concerning angina [222] was criticised on grammatical grounds by London Civil Servants who were planning to administer it. However, the extensive use of the questionnaire in other studies precluded any changes, as comparisons with other populations would then be impossible and further studies on repeatability and validity would have to be carried out on the grammatically improved questionnaire.

16.3.7 The question must be answered

When a question produces embarrassment or offence it may not be answered, even with full guarantees of confidentiality. For example, it may be important to ask questions concerning sexual activity, but older subjects, whether sexually active or not, may be less willing to answer than younger patients. Similarly, religious beliefs may prevent a question's being answered. A Muslim may be as embarrassed if asked whether he drinks alcohol as a Christian would be if asked if he commits adultery.

16.3.8 Should open or closed questions be employed?

An open question is one where the respondent writes his answer in his own words or an interviewer records the exact reply. With a closed question, the respondent chooses between the answer options provided. The question in section 16.2, "Have you any problems?" is an open question and the seven questions in section 16.3 are examples of closed questions. With an open question the subject has to recall something whereas with a closed question he is asked to agree with a statement. This may constitute a fundamental difference between the two varieties of question [223].

Open questions should be used during the initial stages of questionnaire design even when closed questions are finally intended [162]. Mellner considered that the failure to employ open questions may lead to a loss of information [224], but Belson and Duncan suggested that more information may be derived from closed questions [225]. Closed questions certainly minimise the

problem of recall and are most suitable when there are a limited number of possible responses [226]. Responses to open questions may be very difficult to code and analyse [219].

16.3.9 A RESPONSE SET MUST BE DISCOURAGED

A response set is the tendency for a respondent to select a particular answer or to give a certain reply [227]. The individual may tend to select a particular answer whether correct or not: for example, the first answer, the positive response, the negative answer, or the neutral response such as "I don't know." The effect of any response set may be limited by varying the position of any answer options [228] and by providing a greater variety of answers than yes and no [229].

The problem of a response set may tend to be less with open questions and an interviewer than with closed questions and a self-administered questionnaire.

16.4 THE INTERVIEWER-ADMINISTERED QUESTIONNAIRE

There are certain advantages and disadvantages of having the questions read by an interviewer rather than having the respondent complete a self-administered questionnaire (table 16-2). The interviewer must be taught to standardise his or her interview so that a question is always asked in the same manner, in the same order, and after an identical introduction. This training may be difficult and does not ensure that two interviewers will get the same responses. However, interviewers have certain advantages. They will have an impression of whether or not the question is understood and can make additional information available if required. Extra clarifying statements must be stipulated in advance and printed on an interviewer's form. With an interviewer, the sub-

Table 16-2. The advantages of interviewer- and self-administered questionnaires.

Advantages of the interviewer-administered questionnaire	Advantages of self-administered questionnaire
1. When a question is not understood, subsidiary information can be made available to the subject.	1. The conditions of giving the questionnaire can be completely standardised, thus removing observer variability.
2. Subject does not need to be able to read nor have his glasses available.	2. The time taken to train an interviewer and for the interviewer to administer the questionnaire is saved. Therefore the self-administered questionnaire is relatively inexpensive.
3. Multipart questions that depend on an initial positive response are more easily administered.	3. Can be sent through the post, avoiding interviewer's travel and other expenses.
4. Completion rate is higher for the individual questions.	4. May be less embarrassing and answers may be more true.

jects will also be able to answer the questions even if they cannot read or have forgotten their glasses. In addition, when a series of questions has been asked as a consequence of an initial (usually positive) response, the questions are easier to ask using a trained observer, as complicated instructions have to be added to a self-administered questionnaire. For example, the instruction, "If the answer is 'No' please proceed to question . . ." may confuse a respondent. Finally an interviewer can ensure that an answer is obtained to nearly every question, whereas with a self-administered questionnaire some subjects will not answer all the questions. However, the use of an interviewer is expensive and the results subject to observer variation despite rigorous training.

16.5 THE SELF-ADMINISTERED QUESTIONNAIRE

The use of a self-administered questionnaire removes the effect of observer variation. The method also tends to be much less expensive as the questionnaire has only to be handed or posted to the subject and an interviewer does not have to be trained and employed. However, the subjects have to be able to read, and if they can some will not have their reading glasses available when given the self-administered questionnaire and others may fail to complete some of the questionnaire. Where possible, the self-administered questionnaire should be administered under standardised conditions: for example, in a waiting room prior to a clinical investigation.

In a comparison of results from an interviewer and self-administered questionnaire it was found that the self-administered questionnaire gave a higher proportion of positive responses to sensitive questions [159]. For example, male patients were asked about impotence by a male observer, and a self-administered questionnaire including the same questions suggested a higher rate for this complaint (47 percent against 28 percent). It appeared that the patients were reluctant to admit so readily to this embarrassing symptom when asked by an interviewer. In this study, less sensitive questions were not affected by the method of collecting the information. With very personal questions, the self-administered questionnaire may have an advantage.

The self-administered questionnaire is also particularly useful in multicentre international trials where interviewer training and standardised conditions would be difficult to achieve. When there are differences in language between the centres, the questions, whether self-administered or not, must be translated from the original language into the new and then translated back into the original language by someone who has never seen the first questions. If the back translation of a question is not close to the original, the first translator must try again until satisfactory matching is achieved.

Whether a questionnaire is self-administered or not, the origin and purpose of the questionnaire must be fully explained to the respondent and the questions must follow a logical sequence. It must be remembered that the respondent will expect some relationship between adjacent questions; the order of administration may affect the responses [230].

16.6 RANDOMISED CONTROLLED TRIALS IN PSYCHIATRIC PATIENTS

Randomised controlled trials in psychiatric patients are often concerned with subjective changes such as symptoms and may involve greater difficulties when compared with trials in other patients. In psychiatry, new drugs often have to be evaluated directly on patients rather than volunteers, diagnostic difficulties are extreme, and informed consent may be difficult to obtain. The response to treatment may be difficult to measure, treatment may have to be prolonged to exert an effect, habituation may occur to drug treatment, a variable dose may be required, and the drugs in use may produce many side effects.

16.6.1 The assessment of new drugs for use in psychiatry

A new drug for use in cardiovascular medicine may have demonstrable effects in animals and human volunteers and the clinical efficacy of this type of drug can be studied to some extent in these groups. Antiinflammatory and other drugs can be studied in animal models of disease, but animal models of schizophrenia, depression, and anxiety are less well developed. Early in their development, drugs for use in psychiatry have to be tried on patients.

16.6.2 Diagnostic difficulties in psychiatry

Defining a psychiatric condition may present great difficulty. A Medical Research Council Trial of treatments for depression employed an operational definition based on clinical impression, the presence of certain symptoms, a short duration of illness, and lack of previous treatment [231]. However, international differences on whether a mentally disturbed patient has schizophrenia, manic depression, or another condition may exist so that it is very important to carefully define the type of patient who may enter a trial.

16.6.3 Obtaining informed consent from the patient

This problem arises mostly with schizophrenia, severe depression, and mental subnormality. If informed consent cannot be obtained from the patient, it must be obtained from the closest relative or guardian.

16.6.4 Measurement of response

Hamilton [232] has defined four categories of improvement: subjective changes; objective changes; improvement in personal relations; and working capacity.

16.6.4.1 Subjective changes

Symptomatic improvement can be determined by an interviewer or a self-administered questionnaire. When a self-administered questionnaire has been used in psychiatry to diagnose or quantify the degree of depression or anxiety, the questionnaires have been termed *self-rating scales*. Discussion of these scales

is beyond the scope of this book and not all are useful for detecting the response to therapy that may be observed in a trial. For example, the well-known Eysenck Personality Inventory [233] measures whether or not the subject has a neurotic personality. This trait may be constant and not susceptible to short-term fluctuations.

Hamilton has reviewed the self-rating scales that may be employed to assess the state of anxiety. He found that at least one self-administered questionnaire can detect drug effects, namely, Taylor's Manifest Anxiety Scale [234]. Hamilton also considered that three interviewer-administered scales can be useful: the Brief Psychiatric Rating Scale [235] for all psychiatric symptoms; the Symptom Rating Test [236]; and Hamilton's rating scale for anxiety states [237]. A trial in psychiatry must employ standardised methods of assessment, and one or more rating scales should be used where appropriate. The greater the difficulty in assessing a response, the greater the importance of using standardised methods that can be reproduced by another investigator.

16.6.4.2 Changes in personal relationships

Changes in personal relationships may be difficult to assess but the number and duration of contacts with others can be estimated and information sought on the subject's relationships with family, friends, employers, and colleagues at work.

16.6.4.3 Objective changes

The ability to resume full-time work could be an important end point of a trial as could discharge from hospital, discharge from care, readmission to hospital and, in depression, the frequency with which electroconvulsive treatment has to be employed.

Any change in behaviour that can be documented must be carefully defined at the onset of the trial. Hamilton [232] provided two examples: outbursts of temper that could be observed in a hospital ward and the frequency of going out of doors in a patient with a fear of open spaces (agoraphobia).

16.6.5 Problems of drug trials in psychiatric patients

A long duration of treatment may be necessary, the dose of drug may be varied, habituation may occur, and side effects may prove very troublesome. Shepherd [238] considered that the highly lipid soluble drugs that act on the brain may be metabolised at more variable rates than other drugs, making a variable dose schedule of greater importance. Habituation to sedative treatment often results in the side effect of sleepiness being lost after a week or so; therefore, Hamilton suggested starting with very small doses and increasing slowly, even though this will prolong a trial [232]. He also pointed out that a patient habituated to one drug may possibly not respond to a second and cautioned against the use of cross-over trials in psychiatry.

The difficulty of conducting trials in psychiatry should not inhibit such

investigations as even greater problems may arise from a failure to perform randomised controlled trials. The randomised controlled trial is essential in the field of psychiatric investigation.

16.7 THE IMPORTANCE OF MEASURING THE QUALITY OF LIFE DURING A TRIAL

Treatment may produce side effects that interfere with a variety of aspects of the quality of life. For example, antihypertensive medication can produce gastrointestinal effects that interfere with the enjoyment of food and drink, and side effects on the cardiovascular system may reduce sporting activities. Similarly, pharmacological effects may prevent the enjoyment of sex, sedation may interfere with both work and play, and some drugs may produce depression and interfere with personal relationships and social contacts. In a trial, it is not sufficient to demonstrate that an antihypertensive drug lowers blood pressure. It has also to be proved that side effects are not severe and that the quality of life does not deteriorate. It is to be hoped that a treatment improves general well-being and that this effect can be demonstrated.

The factors influencing the quality of life, how this quality can be measured, and the usefulness of such procedures in randomised controlled trials has been discussed recently [239].

16.7.1 Factors affecting the quality of life

The severity of the condition to be treated may affect the quality of life and if the patient is cured an improvement in well-being is to be expected. Unfortunately, certain chronic medical conditions are associated with little in the way of symptoms or disability prior to starting treatment and the latter may result only in an unchanged or reduced quality of life. As discussed earlier, symptomless hypertension is one of these conditions; hyperlipidaemia and mild diabetes mellitus provide other common examples. During a trial of treatment in these conditions any adverse effects of the disease or treatment should be measured. However, it is well recognized that symptomatic complaints are not usually due to the abnormality being treated or the treatment being given but are associated with anxiety and depression [240]. The age, sex, and race of the subject are also associated with the volume of symptomatic complaints [159, 241].

16.7.2 Measurement of the quality of life

An example has been published of the (admittedly retrospective) assessment of the quality of life in a randomised controlled trial [239]—namely, the Veterans Administration trial of treatment for hypertensive patients [37]. In brief, estimates were made of any disablement that would prevent a patient's mobility or ability to work, any disability interfering with other aspects of a subject's life, and any discomfort such as that produced by minor symptom side effects. Table 16-3 lists ten states of well-being and the scores that may be attached to

Table 16-3. Health states or states of well-being (after Fanshel and Bush [242]).

Health state	Score
S_1 Total well-being	1.0
S_2 Minor dissatisfaction Very slight but significant deviation from well-being (e.g., caries, glasses for reading.)	0.975
S_3 Discomfort Subject has a symptomatic complaint with no significant reduction in efficiency.	0.875
S_4 Minor disability Daily activities continue but with a significant reduction in efficiency.	0.8
S_5 Major disability Patients show a severe reduction in efficiency of usual functions.	0.75
S_6 Disabled Unable to go to work but can get about in the community.	0.625
S_7 Confined In an institution.	0.375
S_8 Bedridden	0.125
S_9 Isolated For example, in intensive care.	0.025
S_{10} Comatose	0
S_{11} Dead	0

each state. Following the suggestion of Fanshel and Bush [242], total well-being was arbitrarily allocated a score of one and an eleventh state, death, a score of zero. The scores were based on the assumption that a patient is prepared to trade a certain number of years of life in a reduced state of health for a smaller number of years of life in improved state of health [239, 242]. It must be admitted that the scores were somewhat arbitrary and open to discussion. For example, the scores were calculated on the assumption that a person aged 40 would consider a further 40 years of life in a disabled state to be equal to 25 years of life in a state of total well-being. The outcome for each patient was calculated from the health-status score multiplied by the number of years lived in that state. Thus an overall score was computed for each treatment and termed a *health status index*. As expected, the benefits observed in the actively treated group (an increased survival and reduction in cardiovascular events) proved to be greater than the disadvantages of the symptom side effects and other adverse effects of treatment.

During the course of a randomised controlled trial the health state must be determined by collecting certain information. A subject who has no symptom or disability may be placed in health state S_2 (minor dissatisfaction) as state S_1 can probably only apply to a newborn child. Enquiry should be made into the presence of symptoms and when present, the health state will be reduced to S_3 (discomfort).

A questionnaire may be employed to determine health status during the course of a trial; and possible questions are reproduced in table 16-4. Sixteen questions are given that may prove effective in identifying health states S_3 to

Table 16-4. Sixteen questions that may prove useful in identifying health states S_3 to S_7 during the course of a randomised controlled trial.

Health State	Question
S_3 (Discomfort.)	1. Positive response to any question on symptoms, provided symptom experienced at least once a day on more than half the days under consideration.
S_4 (Minor disability suggested if 2, 3, or 4 is true.)	2. How far can you walk without stopping? (Answer: less than 1 mile.) 3. How many flights of stairs can you climb at one go? (Answer: less than 2.) 4. Has your health or treatment interfered with your hobbies? (Positive answer.)
S_5 (Major disability suggested if 5, 6, 8, or 10 is true.)	5. Answer less than half a mile to Q.2. 6. Answer none to Q.3. 7. Have you been going out to work during the last n months? (Positive answer.) 8. If yes, how many days have you been off sick in the last n months? (Answer: 5 or more days.) 9. Have you been able to do all your usual jobs around the house in the last n months? (Positive answer.) 10. If yes, for how many days in the last n months were you unable to do these jobs through illness? (Answer: 5 or more days.)
S_6 (Disabled state suggested when 11, 12, or 13 is true.)	11. Negative response to Q.7 when subject did not work because of illness. 12. Negative response to Q.9. 13. Can you travel by bus on your own? (Negative answer.)
S_7 (Confined state suggested if 14, 15, or 16 is true.)	14. Are you able to go out and about? (Negative answer.) 15. Do you require assistance with bathing? (Positive answer.) 16. Do you require assistance with dressing? (Positive answer.)

n = number of months.

S_7 . It is hoped that the scores may prove useful in evaluating the results of a randomised controlled trial where one treatment leads to a prolonged life in a poor state of health and a second treatment to a shortened life in a better state of health.

16.8 CONCLUSIONS

In randomised controlled trials attention must not only be directed to objective measurements of outcome but also to the symptomatic well-being of the subjects. This chapter considered how the presence of symptoms should be determined from spontaneous reports and interviewer- and self-administered questionnaires. The characteristics of a good question were discussed includ-

Not suitable for over 65s!
Measuring a mile of climbing stairs not expected
But get it at life undisturbed

ing comprehension, validity, and repeatability. Section 16.6 discussed the importance and difficulty of performing randomised controlled trials of treatment in psychiatric patients where the outcome is often subjective. Finally, section 16.7 extended the assessment of symptoms to other aspects of a patient's life and considered how, in crude terms, the quality of life may be measured.

17. EARLY TRIALS ON NEW DRUGS

This chapter considers the ways in which early trials of new drugs differ from trials of established treatments. Greenwood and Todd [243] have defined three phases of early trials: trials to determine safety and early clinical pharmacology (phase I); trials to determine clinical efficacy and further clinical pharmacology (phase II); and trials for the early clinical development of the drug (phase III). A regulatory authority may be involved in these early trials and provide the approval necessary for the general release or marketing of a drug.

17.1 APPROVAL BY A REGULATORY AUTHORITY FOR EARLY TRIALS

Until recently the usual form of authorisation in the United Kingdom was via a clinical trial certificate. The manufacturer of the new drug applied to the Medicines Division of the Department of Health and Social Security (DHSS) giving the chemistry, pharmacology, and the details of animal experiments with the drug [244]. The DHSS division reported to the Committee on Safety of Medicines (CSM) with advice on whether or not a clinical trial's certificate should be issued.

Outside the United Kingdom, it is often only necessary to inform the regulatory authority of the intention to perform trials on a new substance. No certificate is issued, but the regulatory agency can object. In the United States a notice of Claimed Investigational Exemption for a New Drug is filed with the Food and Drug Administration (FDA). Simon and Jones summarised the countries requiring only notification, and these included most European coun-

tries. They also listed the countries requiring a formal detailed submission and approval by a regulatory authority, for example, Australia, Canada, India, Israel, and South Africa [244].

The manufacturer of a new drug is very concerned with the brain-to-bottle time and the regulations in the United Kingdom led to long delays (sometimes over eight months). The number of clinical trial certificates issued fell from over 170 per year in 1972–1974 to 87 per year in 1980 [245]; early trials were conducted in other countries. In 1981 a new scheme was introduced where exemption from the need to obtain a clinical trial certificate may be granted when the licensing authority receives the following:

1. certified summaries of the basic data
2. a copy of trial protocol
3. confirmation that a medical adviser to the company, working in the United Kingdom, is satisfied that the trial is reasonable.

If the licensing authority objects within 35 days, the pharmaceutical company can still apply for a clinical trial certificate as before [245]. The relaxation of regulations would appear reasonable on two counts: first, exemption has always been possible for doctors and dentists conducting trials on their own initiative and second, phase I trials appear to be very safe.

17.2 PHASE I TRIALS

Phase I trials tend to be open, single-dose studies and, when not a randomised controlled trial, fall outside the scope of this book. However, a randomised controlled trial may be appropriate at this stage. For example, a tranquillizer or antihypertensive drug may be expected to have sedative properties and the dose may be increased stepwise in a trial to determine whether the therapeutic effect occurs at a lower dose than the side effect of sedation. In such a trial, with both subjective and objective assessments, a randomised controlled trial is appropriate with placebo control but only single-blinding. Such a trial differs from studies on established drugs as all adverse effects will be unknown and the trial must be carefully supervised, not double-blind, conducted in a hospital or clinical laboratory, and usually accompanied by haemodynamic and biochemical monitoring. The laboratory must be equipped with all those items that would be required for emergency resuscitation and the staff trained to use this equipment. Written informed consent is, of course, essential.

Phase I studies determine whether the new drug has a pharmacological action that may be useful in treatment, and phase II trials examine whether this action proves a benefit in patients with disease [246].

17.3 PHASE II TRIALS

Phase I studies, having provided data on the safety and clinical pharmacology of the new drug, are extended in phase II in order to determine clinical efficacy

in patients and the doses to be employed. These studies may well be randomised controlled trials, possibly with double-blinding and conducted on outpatients. Careful monitoring for biochemical and other adverse effects will be necessary.

17.4 PHASE III TRIALS

Greenwood and Todd [243] have considered certain objectives for phase III trials: definition of those patients who would benefit from the use of the drug; comparison of the new drug with existing drugs; detection of less common adverse effects; determination of any tolerance to the drug's effect, detection of interactions with other drugs, tobacco, and alcohol; the use of the drug in geriatric and paediatric patients; and further studies on the mode of action. Trials in phase III may differ little from the standard randomised controlled trial.

17.5 REGULATIONS GOVERNING WHETHER A NEW DRUG CAN BE GENERALLY RELEASED

Before the drug regulatory authorities of different countries will authorise the release of a new drug they must be satisfied about the efficacy, safety, and quality control of the product. Norway also requires a medical need for the new drug to be demonstrated. The activities of these authorities have been reviewed by Lumbroso [247].

Stringent clinical trials are now required by all authorities, but in the past countries have varied widely in their requirements. Regulatory authorities were mainly established after therapeutic disasters; for example, in 1937 the Food and Drug Administration (FDA) was created in the United States of America following deaths from a sulphanilamide elixir containing ethylene glycol. In France the regulations were strengthened in 1952 following deaths from a preparation of diethyl tin diiodide. The thalidomide disaster in 1959–1960 led to regulatory authorities being established in many countries. The Committee on Safety of Drugs in the United Kingdom was established in 1964 and involved a voluntary system that became law between 1968 and 1971. Interestingly, Lumbroso points out that in Western Germany (where thalidomide was developed) regulations were first imposed by the EEC in 1972.

The regulatory authorities still vary in the amount and type of information required. One of the most cautious authorities is the FDA, whose deliberations may delay the introduction of new drugs by more than three years. The FDA also requires a copy of all record forms completed during the course of clinical trials. The strict regulations are designed to prevent the introduction of a potentially dangerous drug but it is admitted that, in the process, the public may be deprived of a beneficial drug.

Some countries also require research to be replicated in their country. Lumbroso [247] feared that this requirement could be misused for a commercial

protectionist purpose but pointed out that the replication of studies can bring to light new therapeutic indications and clarify the action of new drugs. However, this additional information may better be sought from different and specially designed studies. The difficulty of devising an internationally acceptable standard trial protocol for new drugs arises from differences in attitudes and legal standards between countries rather than disagreements on scientific merit [248]. Whether a drug will be marketed in a particular country depends not only on the absolute safety in, say, deaths per 10,000 treatments, but also to the country's attitude to such deaths. Where death can follow an infectious disease, an inexpensive antibiotic with some risk is better than a treatment that is safer but too expensive to be provided. However, the promotional activity of pharmaceutical companies may determine which drug is prescribed rather than considerations of cost effectiveness [249]. It would be naive to assume that, at the present, permission to sell a relatively dangerous drug in one country and not another rests on sound humane principles.

17.6 POSTMARKETING TRIALS

Clinical trials on a new drug do not cease with registration for sale. Very important trials may be started after this event in order to examine the long-term efficacy of treatment. Also, prior to registration, it is difficult or impossible to observe a sufficient number of patients for a long enough period to detect rare adverse reactions. It is hoped to overcome this difficulty by monitoring adverse effects in postmarketing surveillance schemes. However, the large randomised controlled trial provides the best opportunity of detecting adverse drug effects as in the trials of clofibrate and oral antidiabetic drugs discussed in chapter 19. Other important randomised controlled trials after marketing include further trials of efficacy in comparison with other drugs, more trials to detect interactions, further trials in certain groups such as the young and old, and more trials to determine the optimum dose and dose frequency. Surprisingly, often little is known about these aspects at the time of registration.

Postmarketing trials have also been termed phase IV studies. One particular variety has been called promotional and has been devised to familiarise clinicians in general with the use of the drug. Such trials must be supported when the drug represents an important new treatment. However, they are usually employed to introduce a further sedative, antiinflammatory, or beta-adrenoceptor blocking drug and the purpose of the trials is to sell the product. The trials are often carried out in general practice and inducements have been provided to persuade a large number of doctors to use the new drug. At the end of such a study the patient may continue to receive the drug at the patient's or government's expense and the pharmaceutical company may recover more than the cost of the trial. An article in the *Sunday Times* of January 29, 1978 summed up the situation.

Patients put at risk as doctors aid drug firms in sales drive

The *Sunday Times* interviewed 39 GP's. Four admitted that they do not tell the patient or ask permission when they are testing a new drug. Twenty-seven said taking part in the trial had influenced their choice of drug, and some said they would never have chosen the drug for the patient if they had not been asked to test it.

This article questioned the ethical position of certain doctors involved in these trials and then criticised the trials for failing to collect important data on adverse events and for paying the doctors to take part. The advantages and disadvantages of promotional trials deserve further attention and their usefulness may depend on the drug being investigated.

Lionel and Herxheimer [250] have stressed that a "good proportion of the drugs available are of little importance in terms of essential health care and are marketed mainly because they can be sold and not because they benefit the health of the population." We should concentrate the limited and valuable resources for randomised controlled trials on fewer drugs.

The clinical investigator should be less willing to investigate a new drug when it is closely similar to many that are already available. These compounds have been termed me-too drugs and lack of interest in them on the part of investigators would help to regulate the provision of such drugs by the pharmaceutical companies. Resources may be better employed in examining more original drugs but it must be admitted that a few me-too drugs prove to have a unique place and to represent a real advance.

17.7 LIMITATIONS OF TRIALS IN NEW DRUGS

Small trials during the early phases of drug development may fail to detect severe adverse reactions and new drugs may be released and subsequently withdrawn from the market. This happened recently with tienilic acid and previously with practolol, even though the latter drug had been used in one large randomised controlled trial involving over 3,000 patients [254].

17.8 CONCLUSIONS

Randomised controlled trials are necessary for the evaluation of new drugs and this chapter considered the type of trial appropriate for each phase in the investigation of a new compound. The contribution of regulatory authorities was discussed as they may have to approve new drugs for use in trials or for general release.

Clinical investigators can be more discriminating in the type of trial in which they become involved. They should avoid promotional trials of me-too drugs which originate from the marketing departments of pharmaceutical companies and should take part in trials of potentially important new drugs or trials of established drugs where an attempt is being made to answer an important question.

18. THE DETECTION OF ADVERSE DRUG REACTIONS

An adverse drug reaction (ADR) has been defined by the World Health Organisation [251] as "one which is noxious, unintended and occurs at doses used in man for prophylaxis, diagnosis or therapy." The detection of symptom side effects was discussed in chapter 16; in this section we shall consider only life-threatening events.

An ADR will be more easily detected when it is an event known to be associated with drug treatment and is relatively rare in the absence of such treatment.

18.1 IS THE EVENT KNOWN TO BE PRODUCED BY DRUGS?

Figure 18-1 gives the hypothetical steps in the recognition of an ADR. If a condition is suspected to be an adverse event and is uncommon in the absence of drug treatment, then an ADR will be detected. If an ADR mimics a common event, usually unassociated with drug treatment, it will be unlikely to be detected. If an ADR is suspected but common in the absence of treatment, it may still be detected as will a previously unsuspected ADR if it is rare in the absence of treatment. The large randomised controlled trial gives the best opportunity for detecting previously unsuspected ADRs provided data on all events are collected during the trial and subsequently analysed.

18.2 FREQUENCY OF THE ADVERSE EVENT

Figure 18-2 provides a schematic representation of the chance of detecting an adverse event during a trial and the alternative, of finding a reduction in this

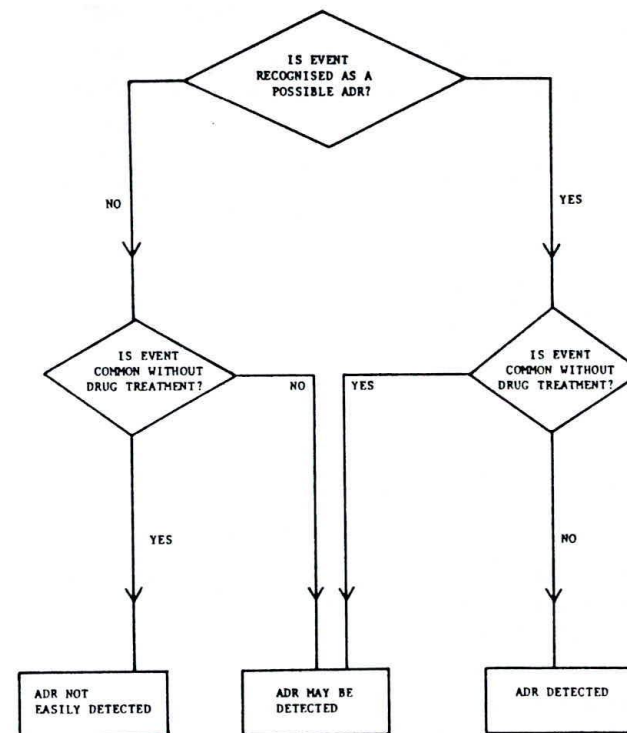


Figure 18-1. Flow chart to illustrate the detection of an adverse drug reaction (ADR) in a trial. Detection is greatest when an event is recognised as a possible ADR and is uncommon in the absence of drug treatment. An ADR may easily be missed when it is not known to be an adverse effect of treatment and is common in the absence of treatment. The large randomised controlled trial provides the best opportunity of detecting such an ADR.

event in the treatment group (a benefit from treatment). An ADR will be detected when the frequency of the event is low in the absence of a particular drug treatment and high when the drug treatment is employed. Conversely, if the drug frequently reduces the incidence of an event that is common without treatment, then a benefit may be detected.

The large randomised controlled trial is the ideal method for detecting ADRs, which occur with a frequency greater than one in 300 treated patients and not in controls [252]. However, trials will never detect rare ADRs such as aplastic anaemia with phenylbutazone (one in 33,000) or thrombocytopenia with diuretics (one in 15,000) [253]. There was also a notable failure of one large trial to detect the more common oculocutaneous syndrome with prazosin [254], possibly due to failure to collect the appropriate information during the trial [255]. With very rare ADRs, the randomised controlled trial is not the appropriate method for detecting these conditions. It is to be hoped that

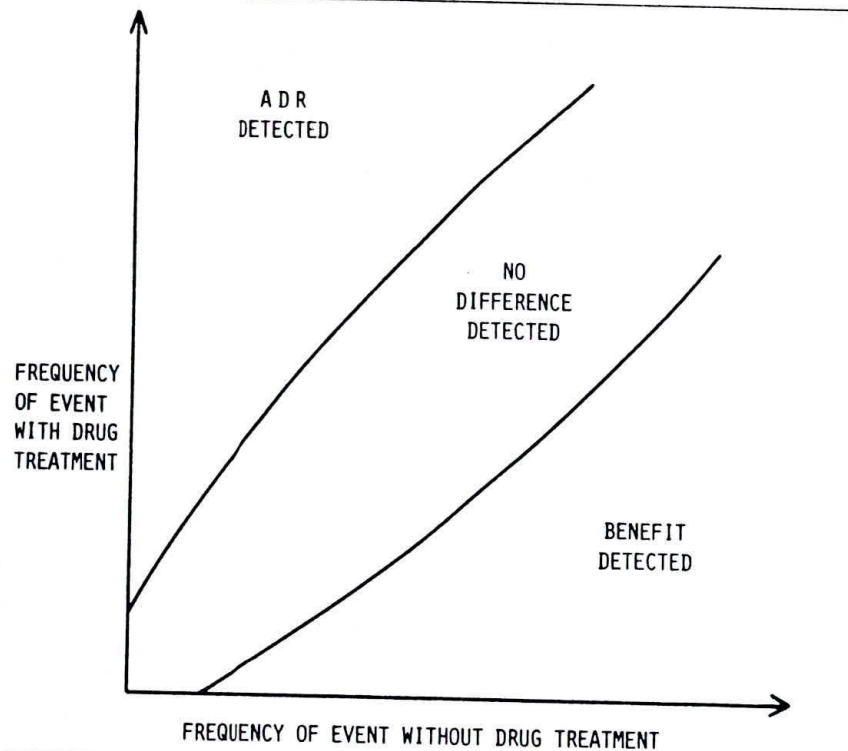


Figure 18-2. Schematic representation of the chance of detecting an adverse drug reaction (ADR) in a randomised controlled trial. When the ADR is common in the treated group and rare in the control group, an ADR may be detected. Conversely, an event that is less common in the treated group will represent a benefit of treatment. The central area represents the results when neither an advantage nor disadvantage of treatment is detected.

clinical intuition, specially designed surveillance programs, and the routine examination of vital statistics will lead to the discovery of rare events. Lewis has prepared a table of the number of patients who would have to be surveyed in a treated group according to the background incidence of an event and the additional incidence due to the drug. With a background incidence of one in 1,000 and a doubling of this incidence due to the drug, 32,000 patients would have to be surveyed [256].

A small clinical trial including less than 50 patients stands little chance of detecting an ADR with any certainty. However, small trials have produced clues to the presence of adverse drug reactions, and large trials have succeeded where the ADR is a common condition. Most life-threatening ADRs occur with a frequency of less than one in 50, an exception being provided by primary pulmonary hypertension with aminorex fumarate which occurred with a frequency of one in ten patients.

18.3 THE SMALL TRIAL AND THE ADR

Adverse drug reactions may be detected in small clinical trials, but the results are often difficult to interpret. Vere discussed this problem [257] and reviewed four trials of chlorpromazine that reported in 1954 when the first case reports of jaundice with this drug were appearing. In these trials 17, 22, 24, and 27 patients were given chlorpromazine. The trial with 27 patients did report one case of jaundice but this could not be attributed definitely to the drug. Similarly, in a trial of the antihypertensive drug guanoxan, two of 160 patients developed obstructive jaundice and one serum jaundice. Seventy percent of the 160 patients showed rises in serum transaminase concentrations during treatment against forty percent before treatment [258]. The authors commented, "Although a rare direct toxic effect . . . upon the liver cannot be excluded . . . this explanation of the remaining abnormalities appears to be most unlikely . . . no case of jaundice has been described which can, with certainty, be ascribed to therapy with these compounds." Many would now consider that the episodes of jaundice were due to the drug.

Trials can also provide evidence of serious toxicity without the end result of the toxic process being observed. For example, a small trial showed evidence of an overall reduction in peak expiratory flow rate with a drug known to produce bronchospasm in susceptible subjects [160]. Similarly, the fact that a drug produces hepatitis may be detected in a small trial in which liver function tests are monitored; and a tendency to produce serious marrow depression may be detected by a significant reduction in platelet or white cell count.

18.4 THE LARGE TRIAL AND THE ADR

Many methods of detecting ADRs depend on case reports, either published or incorporated in central registers. These reports tend to involve drugs that are already under suspicion and also to concentrate on events that have been previously shown to result from drugs—for example, aplastic anaemia and hepatitis. When an adverse effect mimics or is identical to a common disease, it is less likely to be recognised as an ADR. The clinical trial, with its integral control group, is capable of detecting an excess of common events when associated with drug treatment.

Table 18-1 lists some adverse drug effects that have been detected in three large controlled trials. In the Coronary Drug Project Research Group trial [45] there was a one in 50 excess of nonfatal myocardial reinfarctions in those given high-dose conjugated oestrogens for one year, an excess of arrhythmia in those given niacin for a year, and an excess of one in 500 in gallbladder disease in those given clofibrate [259]. In the University Group Diabetes Program trial [46], there was an excess of cardiovascular deaths in patients given tolbutamide and an excess of one in 60 in those given phenformin [260]. In the World Health Organization clofibrate trial [50] there was an excess of total mortality in patients given clofibrate. Oliver has argued that the excess mortality may

Table 18-1. Adverse drug reactions detected in three large controlled trials. The frequency represents the difference between the rate observed in the actively treated group and the appropriate control group.

Drug	ADR	Frequency	Number of patients in trial receiving drug	Reference
CONJUGATED OESTROGEN 5 mg daily for 1 year	Nonfatal myocardial infarction	1 in 50	1,119	Coronary Drug Project [45]
CLOFIBRATE 1.8g daily for 1 year	Gallbladder disease	1 in 500	1,103	Same [259]
NIACIN 3.0g daily for 1 year	Arrhythmia	1 in 100	1,119	Same [259]
TOLBUTAMIDE 1.5g daily for 1 year	Cardiovascular death	1 in 70	204	University Group Diabetes Program [46]
PHENFORMIN 100mg daily for 1 year	Cardiovascular death	1 in 60	204	Same [206]
CLOFIBRATE 1.6g daily for 1 year	Total mortality	1 in 900	5,331	WHO clofibrate trial [50]

not be a direct consequence of the action of clofibrate but could result from the lowering of serum or membrane cholesterol even when achieved by dietary means [261].

The large clinical trial is a very powerful tool for detecting adverse drug effects. However, as mentioned earlier with the practolol trial, a large trial may fail to detect an ADR if the methods of recording such a reaction are not adequate.

18.5 METHODS OF DETECTING AN ADR DURING A TRIAL

The following data should be collected: the fact and cause of any death; the nature of any illnesses; all hospital admissions; and, where indicated, the results of monitoring biochemical and other tests. Determining the fact and cause of death may require the assistance of a National Central Registry such as the Office of Population Censuses and Surveys in the United Kingdom.

18.5.1 Deaths and causes of death

The progress of subjects withdrawn from a trial must be monitored as dropout may be related to an adverse drug reaction or other major end point. The monitoring of defaulters is especially important. A patient in a trial may appear to have defaulted but instead may have died at home or when on holiday. When repeated attempts to contact a defaulter fail, the National Central Registry for Deaths should be consulted to determine whether or not the subject is still alive and, if dead, the cause of death.

18.5.2 Illness, doctor contacts, hospital admissions

Skegg and Doll [255] have argued that the unexpected and unique ocular-cutaneous syndrome associated with practolol may have been detected during a large trial if data on doctor contacts had been collected. There may have been an increased number of consultations with eye problems. A particular ADR may not be anticipated or looked for during the course of a trial but if data are regularly collected on episodes of illness, visits to medical practitioners for any reason, and hospital admissions for any cause, then an excess of a particular episode may be noted when an ADR is occurring. The monitoring of disease episodes should be sensitive enough to detect an increased consultation rate for such items as influenza which may be a drug-induced fever and depression which may be associated with drug administration.

18.5.3 Biochemical monitoring

Where possible, it is obviously preferable to detect only changes in biochemical tests rather than an established ADR. Haematological results, renal function, and hepatic function tests are monitored in most trials. If there is a known probability of an adverse reaction then specific biochemical results can be monitored. For example, a long-term trial in the elderly revealed a deteriora-

tion in glucose tolerance with diuretic treatment before the occurrence of clinical diabetes mellitus [54].

18.5.4 Other investigations

With certain treatments it may be appropriate to monitor other bodily functions: for example, using repeated electrocardiograms, chest radiographs, peak ventilatory flow measurements, or examinations of the optic fundi.

18.6 CONCLUSIONS

The methodology for detecting an ADR in a clinical trial must be improved so that we can be confident that drugs monitored in large long-term trials are safe and to be preferred to those not so evaluated. However, some ADRs will be very difficult to detect, especially changes in affect such as depression or anxiety. How will we determine that a new drug produces schizophrenia in one-half of one percent of cases? The large randomised controlled trial with a full documentation of all disease episodes provides the best hope.

19. FAILURE TO ACCEPT THE RESULTS OF RANDOMISED CONTROLLED TRIALS

The result of a trial may not be accepted for several reasons: the result may be at variance with preconceived ideas; an unusual group of patients may have been recruited; the treatment groups may not be identical in important respects; too few patients may have been recruited and the power of the trial may be too low; the results of the trial may not have been interpreted correctly; the trial result may not be consistent across different strata of patients; the trial may provide a result that conflicts with the results of other trials; the treatment may be difficult to administer or have too many adverse effects; and finally, the trial may originate from a group with a vested interest in demonstrating the observed result (for example, a pharmaceutical company). Before discussing each of these reasons we shall illustrate them by describing three trials whose results have not been completely accepted and also three series of trials on related drugs, the collective results of which are difficult to interpret. After discussing these trials, we shall return to the reasons for rejecting the results of a randomised controlled trial.

19.1 THE ANTURANE REINFARCTION TRIAL

The Anturane Reinfarction Trial was a randomised, double-blind, multicentre trial comparing sulphinpyrazone (Anturane, 200 mg four times a day) with a placebo for the secondary prevention of myocardial infarction. The trial started in September 1975 and an interim analysis in July 1977 revealed a statistically significant ($P \leq 0.02$) reduction in cardiac deaths after an average

Table 19-1. Results of the Anturane (sulphinpyrazone) trial, made available in two reports [48, 242].

	Placebo group	Sulphinpyrazone group	% reduction	P=
First report				
Number patients	742	733	—	
Total mortality (n)	60	40	33	Not specified
All cardiac deaths	?	?	?	—
Cardiac deaths analysed	44	24	49	0.02
Sudden deaths analysed	29	13	57	0.02
Second report				
Number patients	783	775		
Total mortality (n)	85	64	25	Not specified
All cardiac deaths	78	59	24	Not specified
Cardiac deaths analysed	62	43	32	0.06
Sudden deaths analysed	37	22	43	0.04
Sudden deaths (2-7 months)	24	6	74	0.0003
Randomised but ineligible	33	38	—	—
Withdrew from study	220	195	—	—

follow-up period of 8.4 months. The results were published in February 1978 [48] and are given under the heading First Report in table 19-1. Recruitment to the trial was stopped at the time of this report but the investigators disclosed the short-term results to the subjects in the trial and sought their individual consent to continue. All but seven patients agreed to continue; a second report was published in January 1980 [262] and the results are also given in table 19-1.

A 49 percent reduction in cardiac deaths was reported in the first article and a 32 percent decrease in the second. Unexpectedly, the reduction in deaths was not due to the postulated decrease in further episodes of myocardial infarction but to a reduction in sudden deaths, which were possibly related to arrhythmias. The trial has been criticised and the adverse comments concern the definitions employed in the trial, the organisation of the trial, and the manner in which the trial was published.

19.1.1 Definitions employed in the trial

19.1.1.1 Ineligible patients

Ineligible patients were those who were randomised into the trial but were excluded from analysis by the policy committee as the patients did not meet the criteria of the investigation protocol. It appears that some patients in this group were excluded after they had died and therefore the definition of an ineligible patient may crucially affect the results.

19.1.1.2 Nonanalyzable deaths

Nonanalyzable deaths included those occurring either within the first seven days of starting treatment or more than seven days after stopping treatment.

Nonanalyzable deaths therefore included all deaths in patients who dropped out and deaths among patients who did not comply with instructions to take their medication. Nonanalyzable deaths also included those "attributed directly to surgery in which no association could be established with a nonfatal event while the patient was on study treatment" [262]. The nonanalyzable deaths were excluded from an analysis on the per-protocol principle (section 15.7).

19.1.1.3 Sudden death

A sudden death was one that was either not observed or one that occurred within 60 minutes of the onset of symptoms.

19.1.2 Organisation of the trial

The organisation of the trial has been criticised and included the following features:

19.1.2.1 Coordinating centre

The coordinating centre was situated at the Ciba-Geigy Corporation and its Operations Committee was responsible for the execution of trial procedures and the reporting of data to a policy committee. The data were verified by independent university departments of epidemiology and the trial procedures were similarly audited.

19.1.2.2 Policy, Audit, and Electrocardiographic Committees

These committees were independent of the pharmaceutical company.

19.1.2.3 Financing of the trial

The Ciba-Geigy Corporation, the manufacturers of Anturane, financed the trial.

19.1.3 Manner in which the trial was reported

Table 19-1 provides the results of the trial as published in the two reports. There was a fair agreement between the effect of Anturane on sudden deaths; the reduction was 57 percent and 43 percent in the first and second reports, respectively. In the second article it was stated that this reduction all occurred between the second and seventh months of treatment (74 percent reduction in sudden deaths).

19.1.4 Comments that have been published concerning the trial

Four important reviews are summarised in the following sections.

19.1.4.1 Editorial in the Review Epidemiologie et Santé Publique

Armitage wrote this editorial in 1979 [263]. First, he considered the problem of repeated looks (section 10.7.3) and commented, "The ART made no explicit allowance for such repeated testing, but it seems likely that for the particular

design used, whereby the first examination on the data occurred after a substantial proportion of the total patient intake, the adjustment needed would not be great." This statement was made before publication of the second report when, presumably, further looks were undertaken.

When discussing the selection of some end points as either analyzable or sudden deaths (the definitions in 19.1.1) Armitage stated, "The investigators have chosen to discard the safe 'pragmatic' approach in favour of an 'explanatory' approach which may be more sensitive to the presence of real effects but may also suffer from bias."

19.1.4.2 Editorial in the New England Journal of Medicine

In an editorial in January 1980 Braunwald commented [264], "It would certainly be desirable to repeat the sulfinpyrazone study to confirm its results . . ." but " . . . despite the great desirability of learning more about this drug, the information available suggests that sulfinpyrazone should be approved for use after infarction and made available to the American public at the earliest possible time."

The Food and Drug Administration considered whether or not to license the drug for use in the secondary prevention of myocardial infarction and did not agree [265].

19.1.4.3 Article in Science entitled "FDA says No to Anturane"

Kolata [265] reported that the U.S. Food and Drug Administration refused to approve sulfinpyrazone for the secondary prevention of myocardial infarction in April 1980 on the grounds that the case for this drug was not persuasive. She quoted some very important comments. Paul Meier made a remark concerning nonanalyzable deaths: "The idea of nonanalyzable deaths is an innovation in the analysis of clinical trials that we can do without." Meier also commented that the exclusion of ineligible patients magnified the differences between the groups in the trial. In defence, Sol Sherry (chairman of the trial's policy committee) pointed out that bias was prevented by the double-blind nature of the trial and that the exclusion of certain patients was done prospectively and not retrospectively after seeing the data.

Kolata quoted an epidemiologist as saying that the report of the study "was orchestrated [by Ciba-Geigy] for presentation in the scientific and public arena so as to create an impression that there was an unequivocal clear-cut, dramatic result. What happened was almost a con job." Meier was also reported as saying, "It was an interesting but not a convincingly positive result. It was made into a break-through by PR [public relations]."

Most importantly, Kolata reports that Robert Temple, head of the Food and Drug Administration's cardiorenal division, audited the study and found the following:

19.1.4.3.1 THE CLASSIFICATION OF MANY SUDDEN DEATHS WAS INCORRECT. A reclassification removed the deficiency of sudden deaths in the sulfinpy-

razone group. The definition of sudden death "wasn't very well thought out. It turned out to be more crucial than anyone would have anticipated."

It has subsequently been suggested by a member of the trial's policy committee that Temple was biased in his reclassification.

19.1.4.3.2 THE RESULTS WERE NOT GREATLY AFFECTED WHEN NONANALYSABLE DEATHS WERE ALSO INCLUDED IN THE ANALYSIS. A 32 percent reduction in analysable cardiac deaths with the use of sulfinpyrazone was not altered to a great extent and was only lowered to a 24 percent reduction by including nonanalysable deaths.

19.1.4.3.3 THE RESULTS WERE ALTERED WHEN INELIGIBLE PATIENTS WERE INCLUDED. Some patients were ruled ineligible after they had died. Temple considered "everyone in this business knows [such exclusions of dead patients] just are not done" [265].

19.1.4.4 Review in the British Medical Journal

Mitchell pointed out that the numerous exclusion criteria resulted in a highly selected trial population and that the results may not be applicable to other groups of patients with myocardial infarction [266]. He was also worried about the analysis of the trial and considered that the only acceptable analysis of outcome is one based on the intention-to-treat (see section 15.7). Mitchell concluded, "For the present, my verdict on the claim that the report of the ART has altered the state of the art must be 'not proven'."

19.1.5 Conclusions on the Anturane Reinfarction Trial

The basic design of the trial was sound and although the data should have been analysed on both the intention-to-treat and per-protocol basis, I have not seen evidence that an erroneous conclusion was reached. However, anxiety remains over the borderline level of significance achieved and the trial should be repeated. A repeat study would be ethical as substantial proof of benefit is lacking. Section 19.1 illustrates many of the reasons for not accepting the results of trials, and these are summarised in section 19.7.

19.2 A MULTICENTRE TRIAL OF STREPTOKINASE

The European Co-operative Study Group for Streptokinase Treatment conducted a multicentre trial of 24-hour treatment with streptokinase in patients suffering from acute myocardial infarction [267]. Entry to the trial had to be within 12 hours of the onset of chest pain and only 13.5 percent of 2,338 patients with suspected infarction could be entered into the trial. After six months, 48 control patients and 24 treated with streptokinase had died. Table 19-2 gives the deaths according to the time after treatment and reveals the largest benefit from 21 days onwards.

19.2.1 Critical leading article

The results were criticised in a leading article in the *British Medical Journal* [268] for the following reasons:

Table 19-2. Results of the European Co-operative Study Group for Streptokinase Treatment trial according to the time after treatment [267].

Time after treatment	Deaths in the control group	Deaths in the streptokinase group
0-21 days	28	18
21-183 days	20	6

1. The infusion of a lytic agent for the first 24 hours could not influence late mortality, and this was the mortality most affected.
2. Random allocation failed to balance out all the highly relevant risk factors. In the streptokinase group, fewer patients had suffered a myocardial infarction in the past and fewer developed dysrhythmias in the coronary care unit. (The latter may possibly have been due to streptokinase treatment.)
3. There was a low randomisation rate into the trial and many patients were excluded.
4. The leading article went on to consider the merits of the trial and to discuss the possibility that the results were real. However, the trial results were criticised for not fulfilling the author's expectations and producing results that were difficult to implement. To quote, "... the clinical and laboratory complexities inherent in any effective and well-controlled lytic regimen will limit the practical impact of the study on doctors . . ." [268].

19.2.2 Conclusions on the multicentre trial of streptokinase

The criticisms of the multicentre trial do not appear too serious. Low randomisation rates are a common and usually unavoidable problem when early treatment is required, and the assumption that streptokinase could not affect late mortality cannot be supported as acute treatment may limit the size of the infarct. Random allocation often results in groups that are unequal in a small number of respects, but an adjustment can be made during analysis. Finally, although lytic treatment is a difficult procedure there is now a vogue for infusing streptokinase directly into the coronary arteries—a much more onerous task.

19.3 TRIALS OF ASPIRIN IN THE SECONDARY PREVENTION OF MYOCARDIAL INFARCTION

Elwood [269] summarised the results of six randomised controlled trials of aspirin against placebo in the secondary prevention of myocardial infarction. Table 19-3 is derived from his work. Five of the trials demonstrated a reduction in total mortality of between 15 and 30 percent whereas the sixth and largest found an increase in total mortality of 11 percent [275]. However, this large trial (the Aspirin Myocardial Infarction Study) did find a decrease in nonfatal infarctions of 22 percent.

The average reduction in mortality with aspirin was 15 percent (8 percent if

Table 19-3. Summary of total mortality in six randomised controlled trials of aspirin in the secondary prevention of myocardial infarction (derived from Elwood [269]).

Trial	Year	Reference	Total number in trial	Percentage of deaths on aspirin	Percentage of deaths on placebo	Effect of aspirin
Elwood and associates	1974	270	1,239	7.6	9.8	-22%
Coronary drug project	1976	271	1,529	5.8	8.3	-30%
Uberla	1978	272	626	8.5	10.3	-18%
Elwood and Sweetnam	1979	273	1,682	12.3	14.8	-17%
PARIS	1980	274	1,216	10.9	12.8	-15%
AMIS	1980	275	4,524	10.8	9.7	+11%

the result is adjusted according to the numbers in the trials). Aspirin would appear to have a small effect that is difficult to detect in randomised controlled trials.

19.4 TRIALS OF ANTICOAGULANTS IN THE SECONDARY PREVENTION OF MYOCARDIAL INFARCTION

An international anticoagulant review group combined the results of nine trials of long-term anticoagulant administration after myocardial infarction [52]. The pooled results are given in table 19-4 for males and females separately. Total mortality was 20 percent lower in men given anticoagulants ($P < 0.01$) but only 8 percent lower in women. It is easy to understand why anticoagulant therapy was abandoned for women in most countries, but why was this treatment rejected for men? Anticoagulant therapy for myocardial infarction has continued in the Netherlands and is used for both men and women, even the elderly [276]. In most other countries the gains from therapy were not thought to be worth the difficulty of administering anticoagulants. To quote Mitchell [266], "... even if the claims [for anticoagulants] were valid the apparent benefit was too small to justify the hassle of conventional anticoagulant regimens."

A 20 percent reduction in mortality has not been considered worth continuous anticoagulant therapy. This treatment involves repeated estimations of blood coagulability, and bleeding may occur as an adverse effect of too much treatment. The occasional patient may therefore die as a result of treatment and it may be unacceptable to the prescribing clinician to cause one death through treatment even though he may witness several deaths that may (but often may not) have been prevented by treatment.

Of the 16 members of the International Anticoagulant Review Group [52], 14 considered "... the findings warranted a conclusion that anticoagulant therapy probably prolonged survival at least over two years but that benefit was largely restricted to patients with a history of anterior or previous infarction."

The remaining two members "were not convinced that long-term therapy prolonged survival."

19.5 TRIALS OF BETA-ADRENOCEPTOR BLOCKING DRUGS IN THE SECONDARY PREVENTION OF MYOCARDIAL INFARCTION

Vedin [277] gave an address at an international conference and considered the results of five trials of beta-adrenoceptor blocking agents in the secondary prevention of myocardial infarction. Table 19-5 reproduces the results of Vedin's review in which he successively pooled the level of significance for these trials and concluded that beta-adrenoceptor blocking drugs (or at least alprenolol and practolol) were effective in reducing total mortality after a myocardial infarct. Moreover, he suggested that these data were so conclusive that it may be unethical to allow any further placebo-controlled trials on this

Table 19-4. Summary of the results of nine controlled trials of long-term anticoagulant therapy in the secondary prevention of myocardial infarction, prepared by an International Anticoagulant Review Group [52].

	Control group			Anticoagulant group			Reduction in total deaths
	PYE	Deaths	Deaths per 100 PYE	PYE	Deaths	Deaths per 100 PYE	
Males	3,591	340	9.5	3,947	300	7.6	20%
Females	384	43	11.2	476	49	10.2	8%

PYE = person years of exposure.

Table 19-5. Summary of the results of 15 trials using beta-adrenoceptor blocking drugs in the secondary prevention of myocardial infarction. Successively pooled P values according to Vedin [277].

Authors	Reference	Drug	Total numbers		Total deaths		Successively pooled <i>P</i> values	
			Active	Control	Active	Control	Per Compound	TOTAL
RESULTS SELECTED BY VEDIN [277]								
Reynolds and Whitlock	278	Alprenolol	38	39	3	3		
Ahlmark and Saetre	279	Alprenolol	69	93	5	11	0.41	
Wilhelmsen and others	280	Alprenolol	114	116	7	14	0.10	
Andersen and others <65 yrs	281	Alprenolol	140	142	13	29	0.002	
Multicentre International	251	Practolol	1,520	1533	54	83	0.013	0.005

subject. "Nearly 20,000 patients round the world are either already enrolled in or will be enrolled in prospective secondary prevention trials with beta-blockers. This massive program-costing an estimated 30 million dollars a year is unlikely to benefit either patients or science . . ."

Reviewing the same data, Hampton [282] stated, "In 1965, Snow described a clinical trial of propranolol in patients with acute myocardial infarction; he found a considerable reduction in mortality . . . This was the first post-infarction trial of a beta-blocker, and none of the many subsequent trials have demonstrated such marked benefit." Baber and Lewis reviewed the trials on beta-adrenoceptor blocking drugs and published the 90 percent confidence limits [123]. Of 18 trials, eight had confidence limits encompassing a 50 percent increase in mortality, and 14 a decrease in mortality of 50 percent.

Trials now being reported are larger and support the concept that beta-blocking drugs are helpful in the secondary prevention of myocardial infarction [170, 283]. Some of these drugs may be more useful than others in secondary prevention but this would only partly explain the divergent results.

19.6 THE UNIVERSITY GROUP DIABETES PROGRAM TRIAL

The University Group Diabetes Program study was a trial of treatment in newly diagnosed diabetic patients who did not require insulin and had a good prognosis for a five-year survival [46]. Patients were randomly allocated to placebo (PLBO), the oral hypoglycaemic drug tolbutamide (TOLB), a variable dose of insulin (IVAR), or a standard amount of insulin (STD). All patients were given dietary advice. An additional group was randomised to receive phenformin but they were not recruited at the start of the trial and cannot be expected to be identical to the other groups. Table 19-6 gives the results of the trial excluding those in the phenformin group [260]. There was a statistically significant excess of cardiovascular deaths in the group treated with tolbutamide. Total deaths were also increased in this group but the excess did not reach statistical significance. Not surprisingly since tolbutamide is a popular treatment for diabetes, this trial has been criticised.

19.6.1 Criticisms of the University Group Diabetes Program trial

The conclusion that tolbutamide therapy was associated with an excess of cardiovascular deaths has been criticised on the grounds of inadequate data collection, admission of ineligible patients, administration of a fixed dose of tolbutamide, failure to detect a statistical increase in total mortality, inequality of the groups after randomisation, an abnormal outcome in the placebo group, excess deaths not being observed in every clinic, and patients being transferred from one group to another.

19.6.1.1 Inadequate data collection

No data were collected on important risk factors such as smoking or the duration of diabetes prior to entering the trial [284, 285]. There was no way of knowing whether the groups were comparable for these important factors.

Table 19-6. Results of the University Group Diabetes Program (UGDP) Trial.

	Tablet		Insulin	
	Placebo	Tolbutamide	Standard dose	Variable dose
Number	205	204	210	204
Deaths				
Cardiovascular	10	26*	13	12
%	(5)	(13)	(6)	(6)
Non Cardiovascular	11	4	7	6
%	(5)	(2)	(3)	(3)
All causes	21	30	20	18
%	(10)	(15)	(10)	(9)

*P = 0.005 compared with the other groups.

19.6.1.1.1 DEFENCE. Randomisation and the large numbers in the trial make biologically important differences between the groups unlikely.

19.6.1.2 Admission of ineligible patients

Certain patients were admitted to the study who should have been excluded on the basis of a poor prognosis for a five-year survival. These patients may have been unevenly distributed between the groups. On the other hand, in many cases the diabetes was very mild or its presence questionable. These patients should only have been eligible for dietary advice and should not have entered the trial. Sixty-nine patients were admitted without meeting a glucose tolerance test criterion [284, 285].

19.6.1.2.1 DEFENCE. Again, randomisation and the large numbers in the trial would make an important and unequal distribution between the groups unlikely. It can be agreed that patients with borderline diabetes should only receive a diet, but that statement can only be made with confidence now that we know the result of the trial. Also, the dilution of the treatment groups with patients who do not have classical diabetes should not bias the results. However, ineligible patients, as defined by exclusion criteria, should not have been included.

19.6.1.3 Administration of a fixed dose of tolbutamide

A fixed dose of tolbutamide was given to all subjects without regard to their individual needs [285].

19.6.1.3.1 DEFENCE. The result of the trial is relevant to a fixed dose. A variable dose may produce a different result but this remains to be proved.

19.6.1.4 No definite increase in total mortality

There was no statistically significant increase in total mortality in the tolbutamide group and noncardiovascular deaths were reduced in this group.

19.6.1.4.1 DEFENCE. Total mortality was still higher in the tolbutamide group.

19.6.1.5 Inequality of the groups after randomisation

The tolbutamide group had more patients with a high serum cholesterol or major electrocardiographic abnormalities, more males, more obese patients, and more with a history of angina [284]. Similar remarks were made about the group randomised to phenformin [286].

19.6.1.5.1 DEFENCE. The tolbutamide group included fewer hypertensive patients. Moreover, an adjustment for baseline differences did not materially affect the results [287].

19.6.1.6 Abnormal outcome in the placebo group

There were no deaths from myocardial infarction in the placebo group [284].

19.6.1.5.1 DEFENCE. This could have been a chance occurrence because the mortality from myocardial infarction was very low with this treatment

19.6.1.7 The excess deaths were not observed in every clinic

Most of the excess deaths in the group given tolbutamide occurred in only three of the 12 clinics. Schor remarked, "It would appear to any reasonable statistician that for some reason or other the randomisation procedure broke down in these three clinics over some period of time . . ." [284].

19.6.1.7.1 DEFENCE. No evidence has been provided that randomisation broke down. Treatment was assigned by the coordinating centre and not the treating clinic.

19.6.1.8 Transfer of patients from one group to another

Some patients transferred from one group to another and analyses were only performed on the intention-to-treat principle and not by the per-protocol method.

19.6.1.8.1 DEFENCE. There were very few transfers, but both kinds of analysis should have been presented.

19.6.2 Conclusions

The University Group Diabetes Program trial was well designed, and the analysis has not been seriously faulted. When a trial shows a possibly beneficial result of treatment it can be repeated; however, it would not be ethical to repeat a trial to confirm a suspected adverse effect. The results must therefore be accepted at this stage, but as patients are still being given hypoglycaemic drugs, it is hoped that observational studies will clarify any adverse consequences of such treatment and indicate which drugs, if any, may be safely prescribed. It may then be possible to arrange a randomised controlled trial on these compounds. A trial could also be restricted to those patients who do not wish to take insulin and do not diet effectively. These subjects could ethically be randomised to tolbutamide or placebo provided they had symptoms from hyperglycaemia that could benefit from treatment and were also informed of the result of the University Group Diabetes Program trial prior to giving informed consent. This trial would be of possible benefit to the patient and could be justified (section 3.12).

The Committee for the Assessment of Biometric Aspects of Controlled Trials of Hypoglycemic Agents [288] concluded, "We consider that in the light of the UGDP findings it remains with the proponents of the oral hypoglycemics to conduct scientifically adequate studies to justify the continued use of such agents."

19.7 REASONS FOR NONACCEPTANCE OF RANDOMISED CONTROLLED TRIAL RESULTS

The reasons for nonacceptance include: results at variance with preconceived ideas; errors in performance of the trial; errors in analysis; an atypical selection of patients; failure of randomisation to produce equivalent groups; failure to

results within different groups in the same trial and between different trials; the adverse effects of treatment; and the fact that the trial originates from a group with a vested interest in a particular result.

19.7.1 Preconceived ideas not in agreement with the results

The Anturane Reinfarction Trial claimed that Anturane produced a large reduction in sudden deaths after myocardial infarction. This was not expected by most medical practitioners. A reduction was anticipated only for recurrent myocardial infarctions, and the unexpected result was one reason why the result of the Anturane Reinfarction Trial has not been accepted (section 19.1).

The European Trial on Streptokinase treatment showed a reduction in late mortality. This was unexpected and the trial criticised (section 19.2).

19.7.2 Errors in the performance of the trial

Feinstein has criticised the University Group Diabetes Program trial for failing to define terms such as *congestive heart failure*; for having vague selection criteria; for failure to obtain important baseline data and information on the quality of life during the trial; for the quantity of missing data; for the difficulties in standardising the protocol (four clinics initially employed serum rather than whole blood determinations of glucose); and for discontinuing the use of tolbutamide before stopping other treatments [285].

19.7.3 Errors in the analysis of the trial

Errors in analysis were discussed in section 15.1. Three very important errors are discussed next: failure to analyse on the intention-to-treat principle; the effects of repeated looks; and an inadequate classification of end points.

19.7.3.1 Failure to analyse on the intention-to-treat principle

The analysis of the Anturane Reinfarction Trial provides a classic example of the failure to analyse on the intention-to-treat principle (section 15.7, 19.1). Sulphinpyrazone takes seven days to exert a full effect and the effect will be absent seven days after stopping the drug. The investigators therefore excluded patients in whom the drug could not have been active. They also excluded the corresponding placebo treated patients but discarded the safe intention-to-treat approach to analysis in favour of the per-protocol approach which may suffer from bias. Conventionally, patients should be retained in their groups after randomisation [124], and this was the approach adopted by Elwood in his trials of aspirin following myocardial infarction [270, 273].

19.7.3.2 Effect of repeated looks on the significance of the statistical tests reported

It appears that the problem of repeated looks (section 10.7.3) was not considered initially in the Anturane Reinfarction Trial.

19.7.3.3 Classification of end points not well defined

The definition of sudden deaths in the Anturane Reinfarction Trial was inadequate (section 19.1.4), and a reclassification of these deaths may have altered the conclusions to some extent.

19.7.4 Restricted selection of patients at entry

Patients entering a trial do not usually represent patients in general. This fact has led to criticism of many trials of the secondary prevention of myocardial infarction, including the Anturane Reinfarction Trial (19.1.4). Many trials will fail to achieve general validity (chapter 5), but in the Anturane Reinfarction Trial, patients originally diagnosed as having a myocardial infarction were removed from the trial after randomisation, apparently distorting the results (19.1.4). However, the exclusion of patients before randomisation only reduces the general applicability of the results and does not produce bias.

The European trial of streptokinase treatment was criticised as only 13.4 percent of available patients were entered into the trial. Most of the patients were excluded due to inability to give the treatment within 12 hours of the onset of chest pain. It was reasonable to exclude these patients from the trial as the treatment was only thought to exert an effect in the first 12 hours. It would be more sensible to recalculate the inclusion rate with the numbers presenting within 12 hours as the denominator.

19.7.5 Treatment groups not identical at entry

Randomisation worked effectively in the Anturane Reinfarction Trial to give similar groups, but this does not always occur. In the European trial of streptokinase and in the University Group Diabetes Program trial the groups were not identical in some important respects.

19.7.6 Too few patients in the trial

The small numbers of patients in many of the trials of beta-adrenoceptor blocking drugs has led to several being reported as negative but with very low power. Baber and Lewis [123] have graphically illustrated the low power in these trials by providing the 90 percent confidence limits.

A trial may give a negative result not only if too few patients are entered but also if the true effect of treatment is very small. If a treatment confers only a small benefit, the trials have to be larger to prove this with any certainty. For example, if the true reduction in mortality with active treatment is greater than 50 percent, fewer patients will be required to prove this effect than to demonstrate a 20 percent reduction. In the secondary prevention of myocardial infarction, the effects of certain treatments may be small; for example, reductions in total mortality of up to 15 percent for aspirin and 20 percent for anticoagulants.

19.7.7 Faulty interpretation of the trial results

An example of faulty interpretation is provided by trials of antihypertensive agents where baseline blood pressure is determined before the start of the trial. As the patients become accustomed to the procedures adopted during the trial, their average blood pressure tends to fall. This order effect may thus be superimposed on any treatment effect, inflating the apparent effect of treatment. Douglas-Jones and Cruickshank [289] examined the use of atenolol as an antihypertensive agent and compared three different doses in a cross-over, random order, double-blind fashion. Unfortunately, baseline blood pressure was always determined prior to the commencement of the trial. The authors were able to conclude that there was no difference between the blood pressure on the three doses, but they should have been more cautious in concluding, "Atenolol effectively decreased lying and standing blood pressures" as the fall in pressure from the start of the trial may have been enhanced by giving the active doses last. The baseline assessment should have consisted of a double-blind period of placebo treatment given in random order during the body of the trial. However, the authors appear to have determined a correct baseline pressure, neither inflated by observer bias nor order effect, as the result of this trial agrees quantitatively with other trials where baseline blood pressure was determined correctly [290].

19.7.8 Trial results not consistent in different subgroups

It is desirable to examine the results of the trial in subgroups that are obviously important such as the two sexes, different races, and various age groupings. It is less desirable to invent subgroups after examining the data. For example, the best results may be observed in unmarried Chinese women over the age of 70 but the trial may include few such persons and a report on a small selected subgroup may be misleading.

After analysing the results in different groups, the effect of treatment may be shown to be inconsistent; this may raise doubts as to the generality of any conclusions. In an important trial of specialised care for hypertensive patients, the Hypertension Detection and Follow-up Program Trial, small improvements in blood-pressure control were associated with an overall decrease in mortality. The result was not found in white women when analysed separately [132, 211]. This subgroup was not small and raises the question of the generality of the results. Similarly, it appears that men, but not women, may benefit from aspirin in the prevention of venous thromboembolism [291]. In the trials of beta-adrenoceptor blocking drugs, different effects have been reported in the elderly [281] that were not confirmed in other trials [170, 283]. In one trial of these drugs, there was a better response with an anterior myocardial infarction [251] whereas in a second trial a greater treatment effect was observed with a posterior infarct [170]. Great caution has to be employed in subgroup analysis (section 15.7).

19.7.9 Different trials give different results

In three sections on trials in the secondary prevention of myocardial infarction (section 19.3 on the use of aspirin; 19.4 on anticoagulants; and 19.5 on beta-adrenoceptor blocking drugs) different results were apparent, some trials showing benefit and others none. It is a small wonder that the positive results have not been widely accepted. However, combining the results can provide a more clear picture of any overall pattern [52].

19.7.10 The treatment is too difficult or has too many adverse effects

If a treatment is uniformly effective and capable of being provided, it will presumably be offered irrespective of difficulties of administration and laboratory control. However, treatments are not uniformly effective; only a proportion of patients will benefit and the costs and difficulties have to be taken into account. The complexity of treatment was given as a reason for not implementing the results of the European Trial of Streptokinase in acute myocardial infarction (section 19.2).

The trials of anticoagulant treatment following myocardial infarction (section 19.4) indicated a reduction in male mortality of 20 percent, but the complexity of treatment was such that its use has declined except in the Netherlands, where it is still employed.

19.7.11 Vested interest of originating group

The Anturane Reinfarction Trial was criticised on account of being funded and analysed by the pharmaceutical company making Anturane (section 19.1.4), but this trial employed independent policy and audit committees and the criticism should not be taken seriously. However, pharmaceutical companies are responsible for a number of promotional trials and some of these require close attention. These trials are usually concerned with the acceptability of their products and a comparison of these with those of their competitors. The motivation for these trials arises from competitive marketing. An example was provided in section 1.3.

19.8 CONCLUSIONS

This chapter considered 11 reasons why the result of a particular trial may be rejected and gave several examples to support these assertions. As trial design, execution, and analysis improve it is hoped that the proportion of results that are rejected will be reduced. Results will still be falsely positive on occasions, and little can be done to reduce the strength of a reader's preconceived ideas, avoid inconsistency of the treatment effect in different subgroups, or prevent different results emerging from separate trials. However, errors of performance, analysis, and interpretation can be minimised, sufficient representative patients can be recruited, and trials sponsored by the pharmaceutical companies can be monitored and analysed by independent bodies. Even if a result

is accepted it may not lead to any change in clinical practice. A treatment will not be employed if it is considered too difficult, has too many adverse effects, or is too expensive.

As the standard of randomised controlled trials improves, so will the quality of critical appraisal. We must admit that in the future the results of trials may be as frequently rejected as they are today.

20. THE ADVANTAGES AND DISADVANTAGES OF RANDOMISED CONTROLLED TRIALS

This chapter reviews the advantages and disadvantages of performing randomised controlled trials. Observational studies may fail to include controls or may use an inappropriate comparison group. The advantage of the randomised controlled trial lies in the greater confidence with which we can accept the results. Randomised controlled trials have prevented the introduction and continued use of useless and dangerous treatments.

On the debit side, although we may have confidence in a result, the result may nevertheless be incorrect. Also, it is possible that the performance of a randomised controlled trial may delay the introduction of a useful treatment. Further, when a patient is randomised to the treatment that proves least effective, that individual may suffer as a consequence. Although this is true, a randomised controlled trial should only be used when there is genuine doubt about the efficacy of the treatments and the investigator cannot know if patients will suffer. A placebo may not prove to be the less desirable treatment as in some trials this has proven to be the most beneficial therapy [46, 50]. Lastly, it is expensive to perform a randomized controlled trial and when the effect of treatment is very marked it may be unnecessary.

20.1 ADVANTAGES OF RANDOMISED CONTROLLED TRIALS

20.1.1 Randomised controlled trials as the best estimate of a beneficial effect of treatment

The benefit to be derived from treatment must be established by comparing the results with either no treatment or some alternative therapy. Such control

groups may either consist of patients studied in the past or subjects included in a randomised controlled trial where randomisation provides a comparable control group that is treated simultaneously. Attempts to establish that historical controls are adequate [292] have not been supported by the literature [5, 21, 88]. Retrospective data are not valid for comparative purposes if the pattern of disease changes with time or the type of patient differs between two periods of observation. For example, an investigator may study a series of patients and then announce that he intends to evaluate a new treatment in the future. Patients that are referred to him for the new treatment may be more or less severely affected than the control group. However, an anonymous author concluded, "If the change is rapid, . . . , then a randomised trial may be invalidated just as much as any other" [293]. This is difficult to accept as the control group will be treated simultaneously with the treatment under investigation and the result of a randomised controlled trial will be much less influenced by any change than a comparison with historical controls.

20.1.2 Randomised controlled trials as proof that a supposedly beneficial treatment is dangerous

One of the largest contributions to scientific knowledge resulting from the use of randomised controlled trials has been the discovery—not only that a treatment may be useless—but that it may be dangerous. This was discussed in chapter 18, and it cannot be stressed too strongly that even if a treatment appears to have a beneficial action in the short term, the long-term effects must still be assessed in a randomised controlled trial. For example, clofibrate, a drug that lowers serum cholesterol, was compared with placebo in the primary prevention of ischaemic heart disease. Although serum cholesterol was lowered by treatment and some heart attacks were prevented, overall mortality was increased by this drug treatment—an unexpected finding that could not have been detected without a large randomised controlled trial.

20.2 DISADVANTAGES OF RANDOMISED CONTROLLED TRIALS

20.2.1 Falsely negative results

The most common example of a misleading answer is the false negative result. The term *misleading* is used rather than incorrect as a trial provides an estimate of a treatment effect and the authors and readers should consider the confidence limits for that result. A treatment effect may not be statistically significant but if the 95 percent confidence limits lie between a deterioration of 30 percent and an improvement of 50 percent, the conclusion of no proven effect may be misleading but not erroneous. Put another way, if a trial is too small to detect a true benefit, this is not a criticism of randomised controlled trials in general but only of a particular trial. Several small trials may sometimes be combined to give an overall estimate of the effect of a treatment (section 15.8). The methods for combining the results from many trials have

been published [294, 295] and such analyses may be able to utilise the results from small trials.

20.2.2 Falsely positive results

It appears rare for a trial to give a false positive result. At a recent symposium on clinical trials Sir Richard Doll identified one example [296] and Maxwell has quoted another [297]. Maxwell considered that with results achieving the 5 percent level of significance, one in 20 would not be expected to be in error as, "Ethical considerations demand a sound scientific reason for believing that the null hypothesis [no difference between treatments] is certainly not likely to be true. Thus, in clinical research this error is much rarer than 5 percent and its detection even rarer still—also for ethical reasons" [298]. We can agree that a trial with a positive outcome in favour of a treatment is unlikely to be repeated for ethical reasons and therefore a false positive result may not be detected. But a five percent level of significance implies a one in 20 chance of a false positive result when a drug truly has no advantage. However, when the treatment is effective then, by definition, a false positive result will not be observed. A randomised controlled trial is no worse than any other method of investigation in rejecting the null hypothesis when it is correct and the absence of bias in well-conducted randomised controlled trials will reduce the number of false positive results.

When early trials differ in reporting a benefit then a series of trials may be published. Pooling of the results often suggests that the benefit from treatment is small (sections 19.3 and 19.4). Unfortunately, there may be a tendency not to publish the results from trials that prove negative. Both authors and editors will be inhibited from publication and Peto [69] and others [299] have observed that smaller trials tend to show greater positive effects than larger trials. Presumably a large trial, requiring more effort and providing a more dependable result, tends to be published whether negative or otherwise, whereas small trials may be published more frequently when a positive result is observed. Therefore Maxwell has suggested that journal editors "assist the identification of negative results by publishing such work by title only—rather than not at all" [298]. Such papers would have to be reviewed and registered with the editor even if they were not printed.

20.2.3 Delay of a valuable new treatment

An editorial in the *British Medical Journal* discussed early randomised controlled trials with a new treatment and stated [293], "Our current insistence on randomised controlled trials has undoubtedly had a salutary effect on loose thinking but more than once this has been at the expense of progress." Unfortunately, an example was not provided to substantiate the latter claim but it is theoretically possible that the performance of randomised controlled trials will delay the introduction of a new and effective treatment. Fortunately, an insis-

tence on randomised controlled trials may also prevent progress with a useless or dangerous treatment.

20.2.4 Best treatment denied to control group

Fincke discussed a hypothetical trial of a new anticancer drug and concluded that if more patients die in the control group, then the doctors are guilty of manslaughter [300]. However, it can be argued that if a randomised controlled trial is not performed and a new treatment is not discovered to be associated with more deaths than the control treatment, then if the treatment is widely introduced, many doctors will be guilty of manslaughter and many more patients will be victims.

Ethical problems are avoided if there is genuine doubt about the efficacy of a new treatment. Also, when informed consent is sought the patient knows he may be randomised to a control group and gives his consent to take part in the trial. A leading article in the *Lancet* discussed the problems of not obtaining informed consent and quoted a double-blind placebo-controlled trial of the side effects of oral contraceptives. Six pregnancies occurred on placebo when the subjects were not aware that a placebo was being employed, although they were advised to use a spermicidal cream [301, 302]. To avoid ethical problems, the subjects must be fully informed of the nature of the trial. This poses difficulties in paediatrics, psychiatry, and sometimes in the treatment of patients with cancer [303].

When a disease is nearly always fatal and no effective treatment is available, then a randomised controlled trial may be unnecessary since any improvement is obvious [301]. When the outcome is not always predictable a sequential trial or a trial with variable allocation of subjects to treatment may avoid some of the ethical pitfalls of withholding effective treatment from large numbers of patients (section 11.7).

20.2.5 Expense of randomised controlled trials

The performance of a randomised controlled trial is more expensive than uncontrolled observational studies. However, when a comparison group is required to demonstrate a treatment effect, historical controls are not adequate and a randomised controlled trial should be performed. The additional expense will be amply repaid by the improved quality of the data.

20.2.6 Undetectable small treatment effect

Randomised controlled trials do have their limitations and will not detect small effects of treatment. A trial to detect a 10–20 percent reduction in a common event such as myocardial infarction will require an impossibly large number of subjects studied for a long period of time. A large trial may detect a 25–50 percent reduction in a common event but not an equivalent reduction in a rare event.

20.2.7 No detection of rare adverse effects

A very rare adverse effect may not be detected even by a large randomised trial. Clinical impression, monitoring systems, and national vital statistics provide the only hope that a very rare adverse reaction will be discovered [252].

20.2.8 Effect obscured by several confounding factors

An outcome may be affected by many factors other than treatment. Black has argued that if attempts are made to restrict randomisation according to many factors, effective randomisation becomes impossible [304]. The answer is not to restrict randomisation but to allow for any difference retrospectively.

20.3 CONCLUSIONS

This chapter in itself is not intended to persuade the reader that the advantages of randomised controlled trials outweigh the disadvantages. Rather, the entire book is directed to this end. It may be fitting to conclude with Cochrane's finding that, in the Northern Hemisphere, randomised controlled trials have failed to spread to the Catholic South and Communist East. Cochrane considered that the main explanation is the lower extent to which medical students are scientifically educated in these regions. However, he pointed out that Sweden, although situated in the North and West, is an exception. There are few randomised controlled trials carried out in Sweden compared to the number of meticulous observational studies performed. Cochrane admits that other factors may influence the performance of randomised controlled trials: the memory of war crimes in Germany, the extent of private practice and, the authoritarian structure of Soviet medicine [305].

The main advantage of randomised controlled trials lies in the confidence with which we can view the results; the disadvantages are trivial in comparison. I therefore hope that more randomised controlled trials will be performed in the future and their use extended much further in the fields of surgery, obstetrics, orthopaedics, psychiatry, physiotherapy, and sociology. This book is dedicated to any person who embarks on a randomised controlled trial as a result of reading it.

21. REFERENCES

- Bradford Hill, A. The clinical trial. *Br. Med. Bull.* 7:278-282, 1951.
- Cochrane, A.L. *Effectiveness and efficiency: Random reflections on health services*. London, The Nuffield Provincial Hospital Trust, 1972.
- Medical Research Council Investigation. Streptomycin treatment of pulmonary tuberculosis. *Br. Med. J.* 2:769-782, 1948. Material reproduced with permission.
- Doll, R., and Peto, R. Randomised controlled trials and retrospective controls: *Br. Med. J.* 44:280, 1980.
- Rose, G. Bias. *Br. J. Clin. Pharmacol.* 13:157-162, 1982.
- L'Etang, J.C. Historical aspects of drug evaluation. In: *The principles and practice of clinical trials*, Harris, E.L., and Fitzgerald, J.D. (eds.). Edinburgh and London, E. and S. Livingstone, 1970. Pp. 3-6.
- Bull, J.P. The historical development of clinical therapeutic trials. *J. Chronic Dis.* 10:218-248, 1959.
- Bradford Hill, A. *Statistical methods in clinical and preventive medicine*. Edinburgh and London, E. and S. Livingstone, 1962.
- Maitland, C. *Account of inoculating the smallpox*. London, J. Robertson, 1972.
- Williams, W. *Masters of medicine*. London, Pan Books, 1954. P. 23.
- Jenner, E. *An inquiry into the cause and effects of the variolae vaccinae*. London, S. Low, 1798.
- Pearson, G. *An inquiry concerning the history of cowpox*. London, J. Johnson, 1798.
- Waterhouse, B. *A prospect of exterminating the small pox*. Boston, Cambridge, 1800.
- Haygarth, J. *Of the imagination as a cause and cure of disorders of the body*. new ed. Bath, R. Crutwell, 1801. P. 3.
- Sutton, H.G. Cases of rheumatic fever treated for the most part by mint water. *Guy's Hosp. Rep.* 2:392-428, 1865.
- Laplace, P.S. *Théorie analytique des probabilités*. Paris, 1812.
- Louis, P.C.A. *Essay on clinical instruction*. Translated by Martin P. London, S. Highley, 1834.
- Louis, P.C.A. *Recherches sur les effets de la saignée*. Paris, De Mignaret, 1835.
- Bartlett, E. *An essay on the philosophy of medical science*. Philadelphia, Lea and Blanchard, 1844.
- Lister, J. On the effects of the antiseptic system upon the salubrity of a surgical hospital. *Lancet* 1:4 and 40, 1870.
- Pocock, S.J. Randomised controlled trials [letter]. *Br. Med. J.* 1:1661, 1977.
- Vallery-Radot, P. *Pasteur 1822-1895*. Paris, 1922.
- Fibiger, J. Om serumbehandling af Difteri. *Hosp. Tid., Kjøbenhavn* 4:6, 309, 337, 1898.
- Porritt, A.E., and Mitchell, G.A.G. An investigation into the prophylaxis and treatment of wound infection. In: *Penicillin therapy and control in 21 army group*. London, 21 Army Group, 1945. P.7.
- Amberson, J.B., McMahon, B.T., and Pinner, M. Clinical trials of sanocrysin in pulmonary tuberculosis. *Am. Rev. Tuberc.* 24:401-435, 1931.
- Glaser, E.M. Ethical aspects of clinical trials. In: *The principles and practice of clinical trials*, Harris E.L. and Fitzgerald, J.D. (eds.), Edinburgh and London, E. and S. Livingstone, 1970.
- Wade, O.L. Human experiment 2, clinical aspects. In: *Dictionary of medical ethics*, Duncan, A.S., Dunstan, G.R., and Welbourn, R.B. (eds.). London, Darton Longman and Todd, 1977.
- Mitscherlich A., and Mielke, F. *Doctors of infamy: The story of the nazi medical crimes*. New York, Shuman, 1949. Pp. xxiii-xxv.
- World Medical Association. Human experimentation, code of ethics of the World Medical Association. Declaration of Helsinki. *Br. Med. J.* 2:177, 1964.
- Bradford Hill, A. Medical ethics and controlled trials. *Br. Med. J.* 1:1043-1049, 1963. Material reproduced with permission of author and editors of *Br. Med. J.*
- Medical Research Council. Responsibility in investigations on human subjects. *Br. Med. J.* 2: 178-180, 1964.
- Committee Appointed by the Royal College of Physicians of London: Supervision of the Ethics of Clinical Investigations in Institutes. *Br. Med. J.* 2:429-430, 1967. Later report published by the college: *Report of committee on the supervision of the ethics of clinical investigation in institutions*. London, 1973.
- Report of a Committee Appointed by Governor Dwight H. Green of Illinois. Ethics governing the service of prisoners as subjects in medical experiments. *JAMA* 136:457-458, 1948.
- Zelen, M. A new design for randomized clinical trials. *N. Engl. J. Med.* 300:1242-1245, 1979.
- Beecher, H.K. Ethics and clinical research. *N. Engl. J. Med.* 274:1354-1360, 1966.
- Beecher, H.K. Pain, placebos and physicians. *Practitioner* 189:141-155, 1962.
- Veterans Administration Co-operative Study Group on Anti-Hypertensive Agents. Effects of treatment on morbidity in hypertension. II Results in patients with diastolic blood pressure averaging 90 through 114 mm Hg. *JAMA* 213:1143-1152, 1970.
- Beecher, H.K. Surgery as placebo: A quantitative study of bias. *JAMA* 176:1102-1107, 1961. Material reproduced with permission. Copyright 1961, American Medical Association.
- Cobb, L.A., et al. An evaluation of internal-mammary-artery ligation by a double-blind technique. *N. Engl. J. Med.* 260:1115-1118, 1959.
- Dimond, E.G., Kittle, C.F., and Crockett, J.E. Evaluation of internal-mammary artery ligation and sham procedure in angina pectoris. *Circulation* 18:712-713, 1958.
- Adams, R. Internal-mammary-ligation for coronary insufficiency: An evaluation. *N. Engl. J. Med.* 258:113-115, 1958.
- Fish, R.C., Grymes, T.P., and Lovell, M.G. Internal-mammary-artery ligation for angina pectoris: Its failure to produce relief. *N. Engl. J. Med.* 259:418-420, 1958.
- Amery, A., et al. Antihypertensive therapy in elderly patients: Pilot trial of the European Working Party on high blood pressure in the elderly. *Gerontology* 23:426-437, 1977.
- McPherson, K. Statistics: The problem of examining accumulating data more than once. *N. Engl. J. Med.* 290:501-502, 1974.
- Coronary Drug Project Research Group: The Coronary Drug Project. *JAMA* 214:1303-1313, 1970.
- University Group Diabetes Program: A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. *Diabetes* 19, suppl. 2:747-830, 1970.
- Veterans Administration Co-operative Study Group on Anti-hypertensive Agents: Effects of treatment on morbidity in hypertension. Results in patients with diastolic blood pressure averaging 115 through 129 mm Hg. *JAMA* 202:1028-1034, 1967.
- The Anturane Reinfarction Trial Research Group: Sulfapyrazone in the prevention of cardiac death after myocardial infarction. *N. Engl. J. Med.* 298:289-295, 1978.

49. Bulpitt, C.J., Semmence, A., and Whitehead, T. Blood pressure and biochemical risk factors. *Acta Cardiol.* 33:109-110, 1978.
50. Committee of Principal Investigators: A Cooperative trial in the primary prevention of ischaemic heart disease using clofibrate. *Br. Heart J.* 40:1069-1118, 1978.
51. Elwood, P.C., et al. A randomised controlled trial of acetyl salicylic acid in the secondary prevention of mortality from myocardial infarction. *Br. Med. J.* 1:436-440, 1974.
52. An International Anticoagulant Review Group. Collaborative analysis of long-term anticoagulant administration after acute myocardial infarction. *Lancet* 1:203-209, 1970.
53. A progress report of the European Working Party on High Blood Pressure in the Elderly (EWPHE). Cardiac and Renal Function with increasing Age in Elderly Hypertensives. In: *Mild hypertension: Natural history and management*, Gross, F. and Strasser, T. (eds.). Tunbridge Wells, Pitman Medical, 1979. Pp. 181-197.
54. Amery, A., et al. Glucose intolerance during diuretic therapy: Results of trial by the European Working party on Hypertension in the Elderly. *Lancet* 1:681-683, 1978.
55. Fowler, F.G. and Fowler, H.W. *The Pocket Oxford Dictionary of Current English*, SR ed. Oxford Clarendon Press, 1969.
56. Rose, G., and Hamilton, P.J.S. A randomised controlled trial of the effect on middle-aged men of advice to stop smoking. *J. Epidemiol. Community Health* 32:275-281, 1978.
57. Veterans Administration Co-operative Study Group on Antihypertensive Agents. Effects of treatment on morbidity in hypertension: II. Influence of age, diastolic pressure, and prior cardiovascular disease: Further analysis of side effects. *Circulation* 45:991-1004, 1972.
58. Hypertension Detection and Follow-up Program Cooperative Group. Five-year findings of the hypertension detection and follow-up program: II. Mortality by race, sex, and age. *JAMA* 242:2572-2577, 1979.
59. Meier, P. Terminating a trial—The ethical problem. *Clin. Pharmacol. Ther.* 25:633-640, 1979.
60. Stamler, J. When and how to stop a clinical trial: Invited remarks. *Clin. Pharmacol. Ther.* 25:651-654, 1979.
61. Benedict, G.W. LRC Coronary Prevention Trial: Baltimore. *Clin. Pharmacol. Ther.* 25, 685-687, 1979.
62. Schoenberger, J.A. Recruitment to the Coronary Drug Project and the Aspirin Myocardial Infarction Study. *Clin. Pharmacol. Ther.* 25:681-684, 1979.
63. Croke, G. Recruitment for the National Co-operative Gallstone study. *Clin. Pharmacol. Ther.* 25:691-694, 1979.
64. Prout, T.E. Patient recruitment: Other examples of recruitment and solutions. *Clin. Pharmacol. Ther.* 25:695-696, 1979.
65. Zelen, M. The randomisation and stratification of patients to clinical trials. *J. Chron. Dis.* 27:365-375, 1974.
66. Wright, I.S., Marple, C.D., and Beck, D.F. *Myocardial Infarction. Its clinical manifestation and treatment with anticoagulants*. New York, Grune and Stratton, 1954. Pp. 8-10.
67. Weinstein, M.C. Allocation of subjects in medical experiments. *N. Engl. J. Med.* 291:1278-1285, 1974. Material reproduced with permission.
68. Ederer, F. Patient bias, investigator bias and the double-masked procedure in clinical trials. *Am. J. Med.* 58:295-299, 1975.
69. Peto, R. Clinical trial methodology. *Biomedicine Special No.* 24-36, 1978.
70. Peto, R., et al. Design and analysis of randomised clinical trial requiring prolonged observation of each patient. I. Introduction and design. *Br. J. Cancer* 34:585-612, 1976.
71. Roethlisberger, F.G., and Dickson, W.J. *Management and the worker. An account of a research program conducted by the Western Electric Company, Hawthorne Works, Chicago*. Cambridge, Mass., Harvard University Press, 1946.
72. Beecher, H.K. The powerful placebo. *JAMA* 159:1602-1606, 1955.
73. Keats, A.S., Beecher, H.K., and Mosteller, F.C. Measurement of pathological pain in distinction to experimental pain. *J. Appl. Physiol.* 3:34-44, 1950.
74. Keats, A.S., D'Alessandro, G.L. and Beecher, H.K. Report to the council on pharmacy and chemistry. *JAMA* 147:1761-1776, 1951.
75. Beecher, H.K., et al. Field use of methadone and levo-iso-methadone in a combat zone. *U.S. Armed Forces Med. J.* 2:1269-1276, 1951.
76. Lasagne, L., et al. Study of placebo response. *Am. J. Med.* 16:770-779, 1954.
77. Beecher, H.K., et al. Effectiveness of oral analgesics (morphine, codeine, acetylsalicylic acid) and problem of placebo 'reactors' and 'non reactors'. *J. Pharmacol. Exp. Ther.* 109:393-400, 1953.
78. Travell, J., et al. Comparison of effects of alpha-tocopherol and matching placebo on chest pain in patients with heart disease. *Ann. N.Y. Acad. Sci.* 52:345-353, 1949.
79. Evans, W., and Hoyle, C. Comparative value of drugs used in continuous treatment of angina pectoris. *Q. J. Med.* 2:311-338, 1933.
80. Greiner, T., et al. Method for evaluation of effects of drugs on cardiac pain in patients with angina of effort; study of khellin (visamin). *Am. J. Med.* 9:143-155, 1950.
81. Bulpitt, C.J. Heparin as an analgesic in myocardial infarction. A double-blind trial. *Br. Med. J.* 3:279-281, 1967.
82. Jellinek, E.M. Clinical tests on comparative effectiveness of analgesic drugs. *Biomet. Bull.* 2:87-91, 1946.
83. Gravenstein, J.S., Devloo, R.A., and Beecher, H.K. Effect of antitussive agents on experimental and pathological cough in man. *J. Appl. Physiol.* 7:119-139, 1954.
84. Hillis, B.R. The assessment of cough-suppressing drugs. *Lancet* 1:1230-1235, 1952.
85. Shapiro, A.K. Factors contributing to the placebo effect. Their implications for psychotherapy. *Am. J. Psychother.* 18: suppl. 1, 73-88, 1964.
86. Dollery, C.T. A bleak outlook for placebos (and for science). *Eur. J. Clin. Pharmacol.* 15:219-221, 1979.
87. Byar, D.P., et al. Randomised clinical trials. Perspectives on some recent ideas. *N. Engl. J. Med.* 295:74-80, 1976.
88. Christie, D. Before-and-after comparisons: a cautionary role. *Br. Med. J.* 2:1629-1630, 1979.
89. Doll, R., and Peto, R. Mortality in relation to smoking: 20 years' observation on male British doctors. *Br. Med. J.* 2:1525-1536, 1976.
90. Ballintine, E.J. Randomized controlled clinical trial. National Eye Institute workshop for ophthalmologists. Objective measurements and the double-masked procedure. *Am. J. Ophthalmol.* 79:763-767, 1975.
91. Pozdena, R.F. Versuche über Blondlots 'Emission Pesante'. *Ann Physik.* 17:104, 1905.
92. Seabrook, W. *Doctor Wood*. New York, Harcourt, Brace, 1941. P. 234.
93. Fletcher, C.M. Criteria for diagnosis and assessment in clinical trials. In: *Controlled clinical trials*, Hill, A.B. (ed.). Springfield, Ill., Charles C. Thomas, 1960.
94. Kahn, H.A., et al. Serum cholesterol: its distribution and association with dietary and other variables in a survey of 10,000 men. *Isr. J. Med. Sci.* 5:1117-1127, 1969.
95. Wilson, E.B. *An introduction to scientific research*. New York, McGraw-Hill, 1952.
96. Bearman, J.E., Loewenson, R.B., and Gullen, W.H. Muench's postulates, laws, and corollaries or Biometricians' views on clinical studies. *Biometrics* note 4, National Eye Institute, Bethesda, Md., 1974.
97. Foulds, G.A. Clinical research in psychiatry. *J. Ment. Sci.* 104:259, 1958.
98. Report of a co-operative randomised controlled trial. Control of moderately raised blood pressure. *Br. Med. J.* 3:434-436, 1973. Material reproduced with permission.
99. Knowelden, J. In: *Prophylactic trials, medical surveys and clinical trials*, Witts, L.J. (ed.). London, Oxford University Press, 1959.
100. National Diet-Heart Study Research Group. The National Diet-Heart Study. *Circulation* 37, suppl. 1:1253-1259, 1968.
101. Pearson, R.M., Bulpitt, C.J., and Havard, C.W.H. Biochemical and haematological changes induced by tienilic acid combined with propranolol in essential hypertension. *Lancet* 1:697-699, 1979.
102. U.S. Department of Health, Education and Welfare. National Institutes of Health. *Cold study reveals some vitamin C influence; more research needed*. Bethesda, Md. *NIH Record* 25:4, 1973.
103. Heaton-Ward, W.A. Influence and suggestion in a clinical trial (Niamid in mongolism). *J. Ment. Sci.* 108:865-870, 1962.
104. Abraham, H.C., et al. A controlled clinical trial of imipramine (Tofranil) with outpatients. *Br. J. Psychiatry* 109:286-293, 1963.
105. Report to the Medical Research Council by its clinical psychiatry committee: Clinical trial of the treatment of depressive illness. *Br. Med. J.* 1:881-886, 1965.
106. Report of Medical Research Council Working Party on Mild to Moderate Hypertension.

- Randomised controlled trial of treatment for mild hypertension: design and pilot trial. *Br. Med. J.* 1:1437-1440, 1977.
107. Report by the Management Committee. Initial results of the Australian Therapeutic Trial in mild hypertension. *Clin. Sci.* 57:449s-452s, 1979.
108. Cooper, G.R. The World Health Organization Centre for Disease Control Lipid Standardization Program. In: *Quality control in chemistry*. Berlin, Walter de Gruyter and Co., 1976. Pp. 97-105.
109. Rogot, E., and Goldberg, I.D. A proposed index for measuring agreement in test-retest studies. *J. Chron. Dis.* 19:991-1006, 1966.
110. Bulpitt, C.J., Dollery, C.T., and Carne, S. Change in symptoms of hypertensive patients on referral to hospital clinic. *Br. Heart J.* 38:121-128, 1976.
111. Galton, F. Regression towards mediocrity in hereditary stature. *J. Anthropological Inst.* 15:246-263, 1886.
112. Ferris, F.L., and Ederer, F. External monitoring in multiclinic trials: Application from ophthalmologic studies. *Clin. Pharmacol. Ther.* 25:720-723, 1979.
113. Wright, B.M., and Dore, C.F. A random-zero sphygmomanometer. *Lancet* 1:337-338, 1970.
114. Rose, G.A., Holland, W.W., and Crowley, F.A. A sphygmomanometer for epidemiologists. *Lancet* 1:296-300, 1964.
115. Kahn, H.A., et al. Standardising diagnostic procedures. *Am. J. Ophthalmol.* 79:768-775, 1975.
116. Williams, O.D. A framework for the quality assurance of clinical data. *Clin. Pharmacol. Ther.* 25:700-702, 1979.
117. Schwartz, D., Flamant, R., and Lellouch, J. *Clinical trials*. Translated by M.J.R. Healy. London, Academic Press, 1980. Pp. 29-33.
118. Hamilton, M., Thompson, E.M., and Wisniewski, T.K. The role of blood pressure control in preventing complications of hypertension. *Lancet* 1:235-238, 1964.
119. Clark, C.J. and Downie, C.C. A method for the rapid determination of the number of patients to include in a controlled clinical trial. *Lancet* 2:1357-1358, 1966.
120. Coronary Drug Project Research Group. The Coronary Drug Project: Design, methods and baseline results. *Circulation* 47 and 48, suppl. 1:12-137, 1973.
121. National Diet-Heart Study Report. Appendix Aa-c. Sample size estimates for medical trials. *Circulation* 37 and 38, suppl. 1:279-1308, 1968.
122. Rose, G.A. Beta-blockers in immediate treatment of myocardial infarction. *Br. Med. J.* 280:1088, 1980.
123. Baber, N.S., and Lewis, J.A. Beta-blockers in the treatment of myocardial infarction (letter). *Br. Med. J.* 3:59, 1980.
124. Schwartz, D., and Lellouch, J. Explanatory and pragmatic attitudes in therapeutic trials. *J. Chron. Dis.* 20:637-648, 1967.
125. Halperin, M., et al. Sample sizes for medical trials with special reference to long-term therapy. *J. Chron. Dis.* 21:13-24, 1968.
126. George, S.L., and Desu, M.M. Planning the size and duration of a clinical trial studying the time to some critical event. *J. Chron. Dis.* 27:15-24, 1974.
127. Sondik, E.J., Brown, B.W., and Silvers, A. High risk subjects and the cost of large field trials. *J. Chron. Dis.* 27:177-187, 1974.
128. Cochran, W.G. *Sampling techniques*. New York: Wiley, 1963. P. 145.
129. Nam, J.M. Optimum sample sizes for the comparison of the control and treatment. *Biometrics* 29:101-108, 1973.
130. Gail, M., et al. How many controls? *J. Chron. Dis.* 29:723-731, 1976.
131. Dunnett, C.W. Multiple comparison procedure for comparing several treatments with a control. *J. Am. Stat. Assoc.* 50:1096, 1955.
132. Hypertension Detection and Follow-up Program Cooperative Group. Five-year findings of the hypertension detection and follow-up program. 1. Reduction in mortality of persons with high blood pressure including mild hypertension. *JAMA* 242:2562-2571, 1979.
133. Hills, M., and Armitage, P. The two-period cross-over clinical trial. *Br. J. Clin. Pharmacol.* 8:7-20, 1979.
134. Meier, P., and Free, S.M. Further consideration of methodology in studies of pain relief. *Biometrics* 27:576-583, 1961.
135. Fisher, R.A. In: *The design of experiments*. Edinburgh, Oliver and Boyd, 1947.
136. Wilson, C., Pollock, M.R., and Harris, A.D. Diet in the treatment of infective hepatitis. Therapeutic trial of cysteine and variation of fat-content. *Lancet* 1:881-883, 1946.
137. Aenishanslin, W., et al. Antihypertensive therapy with adrenergic beta-receptor blockers and vasodilators. *Eur. J. Clin. Pharmacol.* 4:177-181, 1972.
138. Pearson, R.M., et al. Trial of combination of guanethidine and oxprenolol in hypertension. *Br. Med. J.* 1:933-936, 1976.
139. Chalmers, J., et al. Effects of timolol and hydrochlorothiazide on blood-pressure and plasma renin activity. Double-blind factorial trial. *Lancet* 2:328-331, 1976.
140. Lynch, P., et al. Objective assessment of anti-anginal treatment: a double-blind comparison of propranolol, nifedipine and their combination. *Br. Med. J.* 1:184-187, 1980.
141. Williams, E.F. Experimental designs balanced for the estimation of residual effects of treatments. *Aust. J. Sci. Res. Assoc.* 2:149-168, 1949.
142. Cochran, W.G., and Cox, G.M. In: *Experimental designs*, 2nd ed. New York, Wiley & Sons, 1957. P. 133.
143. Armitage, P. *Sequential medical trials*. Oxford, Blackwell Scientific Publications, 1960.
144. Wald, A. *Sequential analysis*. New York, Wiley, 1947.
145. Robertson, J.D., and Armitage, P. Comparison of two hypotensive agents. *Anaesthesia* 14:53-64, 1959.
146. Snell, E.S., and Armitage, P. Clinical comparison of diamorphine and pholcodine as cough suppressants, by a new method of sequential analysis. *Lancet* 1:860-862, 1957.
147. Anscombe, F.J. Fixed sample-size analysis of sequential observations. *Biometrics* 10:98-100, 1954.
148. Cochran, W.G. Newer statistical methods. In: *Quantitated methods in human pharmacology and therapeutics*, Lawrence, D.R. (ed.). London, Pergamon, 1959. Pp. 119-143.
149. Zelen, M. Play the winner rule and the controlled clinical trial. *J. Am. Stat. Assoc.* 64:131-146, 1969.
150. Meier, P. Terminating a trial—the ethical problem. *Clin. Pharmacol. Ther.* 25:633-640, 1979.
151. Chalmers, T.C. When and how to stop a clinical trial: Invited remarks. *Clin. Pharmacol. Ther.* 25:649-650, 1979.
152. Hill, C., and Sancho-Garnier, H. The two-armed bandit problem, a decision theory approach to clinical trials. *Biomedicine* 28 42-43, 1978.
153. Bearman, J.E. Randomized controlled clinical trial. National Eye Institute Workshop for Ophthalmologists. Writing the protocol for a clinical trial. *Am. J. Ophthalmol.* 79:775-778, 1975.
154. McFate Smith, W. Problems in long-term trials. In: *Mild hypertension: Natural history and management*. Gross, F. and Strasser, T. (eds.). Pitman Medical. Tunbridge Wells, England, 1979. Pp. 244-255.
155. Spriet, A., and Simon, P. Questions à se poser pour vérifier un protocole d'essai thérapeutique avant d'en entreprendre l'exécution. *Thérapie* 32:633-642, 1977.
156. Clinical Trials Unit, Department of Pharmacology and Therapeutics, London Hospital Medical College. Aide-mémoire for preparing clinical trial protocols. *Br. Med. J.* 1:1323-1324, 1977.
157. Hamilton, M. Computer programmes for the medical man: A solution. *Br. Med. J.* 2:1048-1050, 1965.
158. Wright, P., and Haybittle, J. Design of forms for clinical trials (1), (2) and (3). *Br. Med. J.* 2:529-530, 590-592, 650-651, 1979.
159. Bulpitt, C.J., Dollery, C.T., and Carne, S. A symptom questionnaire for hypertensive patients. *J. Chron. Dis.* 27:309-323, 1974.
160. Nicholls, D.P., et al. Comparison of labetalol and propranolol in hypertension. *Br. J. Clin. Pharmacol.* 9:233-237, 1980.
161. Survey Control Unit. Central Statistical Office. *Ask a silly question*. Government Statistical Service HMSO, 1976.
162. Bennett, A.E., and Ritchie, K. *Questionnaires in medicine: A guide to their design and use*. Oxford: Nuff. Prov. Hosp. Trust, 1975.
163. Tinker, M.A. In: *Bases for effective reading*. Minneapolis, Minnesota Press, 1965.
164. Clark, H.H. *Psychol. Rev.* 76:387, 1969.
165. International nonproprietary names (INN) for pharmaceutical substances: WHO, 1976.

166. Bulpitt, C.J., et al. The symptom patterns of treated diabetic patients. *J. Chron. Dis.* 29:571-583, 1976.
167. Bloch, A. *Murphy's law and other reasons why things go wrong*. London, Magnum books, Methuen Paperbacks, 1979.
168. Amery, A., et al. Antihypertensive therapy in patients above age 60: Third interim report of the European Working Party on High Blood Pressure in Elderly (EWPHE). *Acta Cardiol.* 33:113-134, 1978.
169. Hampton, J.R. Presentation and analysis of the results of clinical trials in cardiovascular disease. *Br. Med. J.* 282:1371-1373, 1981.
170. Norwegian Multicentre Study Group. Timolol-induced reduction in mortality and reinfarction in patients surviving acute myocardial infarction. *N. Engl. J. Med.* 301:801-807, 1981.
171. Wilkinson, G.N. Estimation of missing values for the analysis of incomplete data. *Biometrics* 14:257-286, 1958.
172. Gordis, L. Conceptual and methodologic problems in measuring patient compliance. In: *Compliance in health care*, Haynes, R.L., Taylor, R.W., and Sackett, D.L. (eds.). Baltimore and London, John Hopkins University Press, 1979. Pp. 23-65.
173. Sackett, D.L. A compliance practical for the busy practitioner. In: *Compliance in health care*, Haynes, R.L., Taylor, R.W., and Sackett, D.L. (eds.). Baltimore and London, John Hopkins University Press, 1979. Pp. 286-294.
174. Feinstein, A.R., et al. A controlled study of three methods of prophylaxis against streptococcal infection in a population of rheumatic children. II. Results of the first three years of the study, including methods for evaluating the maintenance of oral prophylaxis. *N. Engl. J. Med.* 260:697-702, 1959.
175. Park, L.C., and Lipman, R.S. A comparison of patient dosage deviation reports with pill counts. *Psychopharmacologie* 6:299-302, 1964.
176. Gordis, L., Markowitz, M., and Lilienfeld, A.M. The inaccuracy in using interviews to estimate patient reliability in taking medications at home. *Med. Care* 7:49-54, 1969.
177. Bulpitt, C.J., Clifton, P., and Hoffbrand, B.I. Factors influencing over and underconsumption of anti-hypertensive drugs. *Acta Int. Pharmacodyn Ther. suppl.*:243-250, 1980.
178. Mushlin, A.I., and Appel, F.A. Diagnosing potential noncompliance. Physicians' ability in a behavioural dimension of medical care. *Arch. Intern. Med.* 137:318-321, 1977.
179. Roth, H.P., Caron, H.S., and Hsi, B.P. Measuring intake of a prescribed medication: A bottle count and a tracer technique compared. *Clin. Pharmacol. Ther.* 11:288-337, 1970.
180. Bergman, A.B., and Werner, R.J. Failure of children to receive penicillin by mouth. *N. Engl. J. Med.* 268:1334-1338, 1963.
181. Goldsmith, C.H. The effect of compliance distributors on therapeutic trial. In: *Compliance in health care*, Haynes, R.L., Taylor, R.W., and Sackett, D.L. (eds.). Baltimore and London, John Hopkins University Press, 1979. Pp. 297-308.
182. Feinstein, A.R. Biostatistical problems in 'compliance bias'. *Clin. Pharmacol. Ther.* 16:846-857, 1974.
183. Schor, S., and Karten, I. Statistical evaluation of medical journal manuscripts. *JAMA* 195:1123-1128, 1966.
184. Schoolman, H.M., et al. Statistics in medical research: principles versus practices. *J. Lab. Clin. Med.* 7:357-367, 1968.
185. Lionel, N.D.W., and Herxheimer, A. Assessing reports of therapeutic trials. *Br. Med. J.* 3:637-640, 1970.
186. Gore, S.M., Jones, I.G., and Rytter, E.C. Misuse of statistical methods: critical assessment of articles in B.M.J. from January to March 1976. *Br. Med. J.* 1:85-87, 1977.
187. Freiman, J.A., et al. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 'negative' trials. *N. Engl. J. Med.* 299:690-694, 1978.
188. Glantz, S.A. Biostatistics: How to detect, correct and prevent errors in the medical literature. *Circulation* 61:1-7, 1980.
189. Armitage, P. *Statistical methods in medical research*. Oxford and Edinburgh, Blackwell Scientific, 1971.
190. Snedecor, G.W., and Cochran, W.G. *Statistical methods*. Ames, Iowa, Iowa State University Press, 1967.
191. Petrie, A. *Lecture notes on medical statistics*. Oxford and Edinburgh, Blackwell Scientific, 1978.
192. Peto, R., et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II: Analysis and examples. *Br. J. Cancer* 35:1-39, 1977.
193. Hogben, L., and Sim, M. The self-controlled and self-recorded clinical trial for low-grade morbidity. *Br. J. Prev. Soc. Med.* 7:163-179, 1953.
194. Bulpitt, C.J. The design of clinical trials. *Br. J. Hosp. Med.* 13:611-620, 1975.
195. Feinstein, A.R. Clinical biostatistics. A survey of the statistical procedures in general medical journals. *Clin. Pharmacol. Ther.* 15:97-107, 1974.
196. Zar, J.H. In: *Biostatistical analysis*. Englewood Cliffs, N.J., Prentice-Hall, 1974. Pp. 130-131.
197. Nie, N.N., et al. *Statistical package for the social sciences*. New York, McGraw-Hill, 1975.
198. Healy, M.J.R., and Whitehead, T.P. Outlying values in the national quality control scheme. *Ann. Clin. Biochem.* 17:78-81, 1980.
199. Barnett, V. The study of outliers: Purpose and model. *Appl. Stat.* 27:242-250, 1978.
200. John, J.A. Outliers in factorial experiments. *Appl. Stat.* 27:111-119, 1978.
201. Sprackling, M.E., et al. Blood pressure reduction in the elderly: a randomised controlled trial of methylodopa. *Br. Med. J.* 283:1151-1153, 1981.
202. Mantel, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.* 50:163-170, 1966.
203. Breslow, N.E. Analysis of survival data under the proportional hazards model. *Int. Stat. Rev.* 43:45-57, 1975.
204. Wulff, H.R. Letter: Confidence limits in evaluating controlled therapeutic trials. *Lancet* 2:969-970, 1973.
205. Leading article. Interpreting clinical trials. *Br. Med. J.* 2:1318, 1978.
206. Baber, N.S., and Lewis, J.A. Beta-blockers in treatment of myocardial infarction. *Br. Med. J.* 2:59, 1980.
207. Gore, S.M. Statistics in question: Assessing methods—confidence intervals. *Br. Med. J.* 283:660-662, 1981.
208. Ederer, F. A parametric estimate of the standard error of the survival rate. *J. Am. Stat. Assoc.* 56:111-118, 1961.
209. Cox, D.R. Regression models and life tables. *J. R. Stat. Soc. Br.* 34:187-202, 1972.
210. McMichael, J. Anticoagulants: Another view. *Br. Med. J.* 2:1007, 1964.
211. Hypertension Detection and Follow-up Program Cooperative Group. Five-year findings of the hypertension detection and follow-up program. II. Mortality by race-sex and age. *JAMA* 242:2572-2577, 1979.
212. Amery, A., et al. Hypotensive action and side effects of clonidine-chlorthalidone and methylodopa-chlorthalidone in treatment of hypertension. *Br. Med. J.* 4:392-395, 1970.
213. McMahon, F.G. Efficacy of an antihypertensive agent. Comparison of methylodopa and hydrochlorothiazide in combination and singly. *JAMA* 231:155-158, 1975.
214. Gibb, W.E., et al. Comparison of bethanidine, alpha-methylodopa and reserpine in essential hypertension. *Lancet* 2:275-277, 1970.
215. Hefferman, A., et al. A within-patient comparison of debrisoquine and methylodopa in hypertension. *Br. Med. J.* 1:75-78, 1971.
216. Conolly, M.E., et al. A crossover comparison of clonidine and methylodopa in hypertension. *Eur. J. Clin. Pharmacol.* 4:222-227, 1972.
217. Oates, J.A., et al. The relative efficacy of guanethidine, methylodopa and pargyline as antihypertensive agents. *N. Engl. J. Med.* 273:729-734, 1965.
218. Prichard, B.N.C., et al. Bethanidine, guanethidine and methylodopa in treatment of hypertension: A within patient comparison. *Br. Med. J.* 1:135-144, 1968.
219. Schooler, K.K. A study of errors and bias in coding responses to open end questions. *Dissertation Abstr.* 16:2542, 1956.
220. Young, D.W. Evaluation of a questionnaire. *Methods Inf. Med.* 11:15-19, 1972.
221. Collen, M.F., et al. Reliability of a self-administered medical questionnaire. *Arch. Intern. Med.* 123:664-681, 1969.
222. Rose, G.A. The diagnosis of ischaemic heart pain and intermittent claudication in field surveys. *Bull. WHO* 27:645-658, 1962.
223. Maccoby, E.E., and Maccoby, N. The interview: A tool of social science. In: *Handbook of social psychology*, Lindzey, G. (ed.). Reading, Mass., Addison-Wesley, 1954.
224. Mellner, C. The self-administered medical history. Theoretical possibilities and practical limitations of the usefulness of standardized medical histories. *Acta Chir. Scand. Suppl.* 406:1+, 1970.

225. Belson, W., and Duncan, J.A. A comparison of the check-list and the open response questioning systems. *Appl. Stat.* 11:120-132, 1962.
226. Cannel, C.F., and Kahn, R.L. The collection of data by interviewing. In: *Research Methods in Behavioral Sciences*, Festinger, L., and Katz, D. (eds.). New York, Dryden, 1953.
227. Cronbach, L.J. Further evidence on response sets and test design. *Educ. Psych. Meas.* 10:3, 1950.
228. Anderson, J., and Day, J.L. New self-administered medical questionnaire. *Br. Med. J.* 4:636-638, 1968.
229. Anastasi, A. In: *Psychological testing*. New York, Macmillan, 1961.
230. Parten, M.B. In: *Surveys, polls and samples*. New York, Harper, 1950.
231. Shepherd, M. Implications of a multi-centred clinical trial of treatment of depressive illness. In: *Anti-depressant drugs*, Garattini, S., and Dukes, M.N.G. (eds.). Amsterdam, Excerpta Medica, 1967. P. 332.
232. Hamilton, M. Evaluation of psychotropic drugs (3) sedatives. In: *Principles and practice of clinical trials*, Harris, E.L., Fitzgerald, J.D., (eds.). Edinburgh and London, E. and S. Livingston, 1970. Pp. 217-225.
233. Eysenck, H.J. *Manual of the Maudsley personality inventory*. London: University of London Press, 1959.
234. Taylor, J.A. A personality scale of manifest anxiety; *J. Abnorm. Soc. Psychol.* 48:285, 1953.
235. Overall, J.E., and Gorham, G.R. The brief psychiatric rating scale. *Psychol. Rep.* 10:799, 1962.
236. Kellner, R., and Sheffield, B.F. The use of self-rating in a single-patient multiple cross-over trial. *Br. J. Psychiat.* 114:193-196, 1968.
237. Hamilton, M. The assessment of anxiety states by rating. *Br. J. Med. Psychol.* 32:50, 1959.
238. Shepherd, M. Evaluation of psychotropic drugs. In: *Principles and practice of clinical trials*, Harris, E.L., and Fitzgerald, J.D. (eds.). Edinburgh and London, E. and S. Livingston, 1970. Pp. 208-216.
239. Bulpitt, C.J. Quality of life in hypertensive patients. In: *Hypertensive cardiovascular disease: Pathophysiology and treatment*, Amery, A. et al. (eds.). The Hague, Martinus Nijhoff, 1982.
240. Bulpitt, C.J., Dollery, C.T., and Hoffbrand, B.I. The contribution of psychological features to the symptoms of treated hypertensive patients. *Psychol. Med.* 7:661-665, 1977.
241. Bulpitt, C.J., et al. The symptom patterns of treated diabetic patients. *J. Chron. Dis.* 29:571-583, 1976.
242. Fanshel, S., and Bush, J.W. A health status index and its application to health services outcomes. *Operations Res.* 18:1021-1065, 1970.
243. Greenwood, D.T., and Todd, A.H. From laboratory to clinical use. In: *Clinical trials*, Johnson, F.N., and Johnson, S. (eds.). Oxford, Blackwell Scientific Publications, 1977, Pp. 13-35.
244. Simon, T.R.M., and Jones, G. Safety of medicines: The control of clinical trials. In: *Clinical trials*, Johnson, F.N., and Johnson, S. (eds.). Oxford, Blackwell Scientific Publications, 1977. Pp. 1-2.
245. Griffin, J.P., and Long, J.R. New procedure affecting the conduct of clinical trials in the United Kingdom. *Br. Med. J.* 283:477-478, 1981.
246. Dollery, C.T. Clinical trials of new drugs. *J. Roy. Coll. Phys.* 11:226-233, 1977.
247. Lumbroso, A. The introduction of new drugs. In: *Pharmaceuticals and health policy*, Blum, R., et al. (eds.). London, Croom Helm, 1981.
248. Crout, J.R. Quoted in Lumbroso, A. The introduction of new drugs. In: *Pharmaceuticals and health policy*, Blum, R., et al. (eds.). London, Croom Helm, 1981.
249. Silverman, M., and Lydecker, M. The promotion of prescription drugs and other puzzles. In: *Pharmaceuticals and health policy*, Blum, R., et al. (eds.). London, Croom Helm, 1981. Pp. 78-92.
250. Lionel, W., and Herxheimer, A. Coherent policies on drugs: Formulation and implementation. In: *Pharmaceuticals and health policy*, Blum, R., et al. (eds.). London, Croom Helm, 1981. P. 240.
251. WHO. *Tech. Rep. Ser.* 425:5, 1969.
252. Bulpitt, C.J. Screening for adverse drug reactions. *Br. J. Hosp. Med.* 18:329-334, 1977.
253. Böttiger, L.E., and Westerholm, B. Drug-induced blood dyscrasias in Sweden. *Br. Med. J.* 3:339-343, 1973.
254. Multicentre international study: Reduction in mortality after myocardial infarction with long-term beta-adrenoceptor blockade. Supplementary Report. *Br. Med. J.* 2:419-421, 1977.
255. Skegg, D.C.G., and Doll, R. The case for recording events in clinical trials. *Br. Med. J.* 2:1523-1524, 1977.
256. Lewis, J.A. Post marketing surveillance: How many patients? *Tips* (April):93-94, 1981.
257. Vere, D.W. Controlled trials to detect efficacy and toxicity: Training to meet tomorrow's needs. In: *Principles and practice of clinical trials*, Harris, E.L., and Fitzgerald, J.D. (eds.). Edinburgh and London, E. and S. Livingston, 1970. Pp. 242-249.
258. Mann, R.D., et al. The significance of variations in the serum transaminases in the assessment of two new drugs. In: *Experimental studies and clinical experience—the assessment of risk*, Proc. Eur. Soc. for the Study of Drug Toxicity, vol. VI. Amsterdam, Excerpta Medica Foundation, 1965.
259. The coronary Drug Project Research Group. Clofibrate and niacin in coronary heart disease. *JAMA* 231:360-381, 1975.
260. Knatterud, G.L., et al. Effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. IV. A preliminary report on phenformin results. *JAMA* 217:777-784, 1971.
261. Oliver, M.F. Serum chloesterol—the knave of hearts and the joker. *Lancet* 2:1090-1095, 1981.
262. The Anturane Reinfarction Trial Research Group: Sulfinpyrazone in the prevention of sudden death after myocardial infarction. *N. Engl. J. Med.* 302:250-256, 1980.
263. Armitage, P. Trials of antiplatelet drugs: some methodological considerations. *Rev. Epidemiol. Sante Publique* 27:87-90, 1979.
264. Braunwald, E. Treatment of the patient after myocardial infarction. The last decade and the next. *N. Engl. J. Med.* 302:290-292, 1980.
265. Kolata, G.B. FDA says no to Anturane. *Science* 208:1130-1132, 1980.
266. Mitchell, J.R.A. Secondary prevention of myocardial infarction. The present state of the ART. *Br. Med. J.* 2:1128-1130, 1980.
267. European Cooperative Study Group for Streptokinase Treatment in Acute Myocardial Infarction. Streptokinase in acute myocardial infarction. *N. Engl. J. Med.* 301:797-802, 1979.
268. Leading Article. Fibrinolytic therapy in myocardial infarction. *Br. Med. J.* 2:1017-1018, 1979.
269. Elwood, P.C. Aspirin, dipyridamole and secondary prevention. In: *The clinical impact of beta-adrenoceptor blockade*, Burley, D.M., and Birdwood, G.F.B. (eds.). Horsham, England, Ciba Laboratories, 1980. Pp. 25-26.
270. Elwood, P.C., et al. A randomised controlled trial of acetyl salicylic acid in the secondary prevention of mortality from myocardial infarction. *Br. Med. J.* 1:436-440, 1974.
271. Coronary Drug Project Research Group: Aspirin in coronary disease. *J. Chron. Dis.* 29:625-642, 1976.
272. Uberla, K. In: *Acetylsalicylic acid in cerebral ischaemia and coronary heart disease. IV. Colfarit Symposium Berlin 1977*. Stuttgart and New York, Schattauer Verlag, 1978. P. 157.
273. Elwood, P.C., and Sweetnam, P.M. Aspirin and secondary mortality after myocardial infarction. *Lancet* 2:1313-1315, 1979.
274. Persantine and aspirin in coronary heart disease. The Persantine-Aspirin Reinfarction Study Research Group. *Circulation* 62:449-461, 1980.
275. Aspirin Myocardial Infarction Study Group: A randomized, controlled trial of aspirin in persons recovered from myocardial infarction. *JAMA* 243:661-669, 1980.
276. Sixty Plus Reinfarction Study Research Group: A double-blind trial to assess long-term oral anticoagulant therapy in elderly patients after myocardial infarction. *Lancet* 2:989-993, 1980.
277. Vedin, J.A. Analysis presented at the VIIIth European Congress of Cardiology. Paris, 1980.
278. Reynolds, J.L., and Whitlock, R.M. Effects of beta-adrenergic receptor blocker in myocardial infarction for one year from onset. *Br. Heart J.* 34:252-259, 1972.
279. Ahlmark, G., and Saetre, H. Long-term treatment with β -blocker after myocardial infarction. *Eur. J. Clin. Pharmacol.* 10:77-83, 1976.
280. Wilhelmsson, C., et al. Reduction of sudden deaths after myocardial infarction by treatment with alprenolol. Preliminary results. *Lancet* 2:1157-1160, 1974.
281. Andersen, M.P., et al. Effect of alprenolol on mortality among patients with definite or suspected acute myocardial infarction. Preliminary results. *Lancet* 1:865-868, 1979.

282. Hampton, J.R. Evidence on use of beta-blockers. In: *Current themes in cardiology*, Birdwood, G.F.B., and Russel, J.G. (eds.). Horsham, England, Geigy Pharmaceuticals, 1981. Pp. 86-89.
283. Beta blocker heart attack study group. The beta blocker heart attack trial. *JAMA* 246:2073-2074, 1981.
284. Schor, S. The university group diabetes program. A statistician looks at the mortality results. *JAMA* 217:1671-1675, 1971.
285. Feinstein, A.R. An analytic appraisal of the university group diabetes program (UGDP) study. *Clin. Pharmacol. Ther.* 12:167-191, 1971.
286. Data has holes, statistical men report to USV. *Drug Trade News*, 23 August 1971.
287. Cornfield, J. The University Group Diabetes Program. A further statistical analysis of the mortality findings. *JAMA* 217:1676-1687, 1971.
288. Committee for the assessment of biometric aspects of controlled trials of hypoglycemic agents. *JAMA* 231:585-608, 1975.
289. Douglas-Jones, A.P., and Cruickshank, J.M. Once daily dosing with atenolol in patients with mild or moderate hypertension. *Br. Med. J.* 1:990-991, 1976.
290. Bulpitt, C.J., and Rose, G.A. Design and conduct of clinical trials. In: *The hypertensive patient*. Marshall, A.J., and Barritt, D.W. (eds.). Tunbridge Wells, Pitman Medical, 1980. Pp. 494-505.
291. Harris, W.H. et al. Aspirin prophylaxis of venous thromboembolus after total hip replacement. *N. Engl. J. Med.* 297:1246-1249, 1977.
292. Cranberg, L. Do retrospective controls make clinical trials 'inherently fallacious'? *Br. Med. J.* 2:1265-1266, 1979.
293. Leading article. Randomised controlled trials? *Br. Med. J.* 4:1244-1245, 1979.
294. Woolf, B. On estimating the relationship between blood group disease. *Ann. Hum. Genet.* 19:251-253, 1955.
295. Mantel, N., and Haenszel, W. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* 22:719-748, 1959.
296. Mitra, B. Potassium, glucose and insulin in treatment of myocardial infarction. *Lancet* 2:607-609, 1965.
297. Hargreaves, M.A., and Maxwell, C. The speed of action of desipramine: a controlled trial. *Int. J. Neuropsychiat.* 3:140-141, 1961.
298. Maxwell, C. Clinical trials, reviews, and the journal of negative results. *Br. J. Clin. Pharmacol.* 1:15-18, 1981.
299. Rogers, S.C., and Clay, P.M. A statistical review of controlled trials of imipramine and placebo in the treatment of depressive illness. *Br. J. Psychiat.* 127:599-603, 1975.
300. Fincke, M. Arzneimittelprüfung: Strafbare Versuchsmethoden. Heidelberg/Karlsruhe, 1977.
301. Leading Article. Controlled trials: Planned deception? *Lancet* 1:534-535, 1979.
302. Goldzieher, J.W. et al. A double-blind cross-over investigation of the side effects attributed to oral contraceptives. *Fertil. Steril.* 22:609-623, 1971.
303. Pryce, I.G. Clinical research upon mentally ill subjects who cannot give informed consent. *Br. J. Psychiat.* 133:366-369, 1978.
304. Black, D. The paradox of medical care. *J. Roy. Coll. Phys. Lond.* 13:57-65, 1979.
305. Cochrane, A.L. Attitudes to controlled trials. Paper presented at the 9th International Scientific Meeting of the IEA. Edinburgh, August 1981.

INDEX

- Abraham HC, 73
- accepted treatment, 2
- accuracy of a measurement, 81-83
- Adams R, 23
- adverse drug reactions
 - detection, 26-27, 214-220 (see detection of an ADR)
 - recompensation, 13
 - terminating a trial, 26
- Aenishanslin W, 125
- Ahlmark G, 229
- Allessandro GRD, 57
- allocation of treatment
 - according to date, 54
 - hospital number, 54
 - name, 54
 - preceding results, 54
 - wishes of patient, 48, 54
- free of bias, 45-46
 - random (see randomisation)
- Amberson JB, 10
- American Heart Association Anticoagulants Trial, 45-46
- Amery A, 23, 34, 162-163, 166, 176, 196, 220
- analysis
 - amalgamating trial results, 192-193
 - experimental unit, 180-181
 - errors, 180-185
 - for effective randomisation, 185
 - intention-to-treat, 191-192
 - lumping and splitting, 191-192
 - meaning of P, 183-184
 - nonparametric, 186
 - no statistical test, 181
 - per-protocol, 192
 - proportional data, 187
 - quantitative data, 185
 - survival data 187-189
 - wrong statistical test, 181-182
- Anastasi A, 201
- Anderson J, 201
- Andersen MP, 229
- angina pectoris, 21
- ANOVA, 197
- Anscombe FJ, 132
- anthrax, 10
- anticoagulants, 2, 31-32, 45-46, 78, 218, 228
- antidepressant drugs, 63-64
- antihypertensive treatment, 22-26, 30, 36-37, 74-75, 76-78, 102-104, 112-113, 121, 160-162, 165-166, 205, 236
- Anturane Reinfarction Trial, 26, 31-32, 64, 221-225
- Appel FA, 171
- Armitage P, 5, 120, 128-131, 179, 183, 185-187, 223-224
- aspirin, 31-32, 226-227

- Aspirin Myocardial Infarction Study, 40–42, 226–227
- Australian Therapeutic Trial in mild hypertension, 77
- Baber NS, 104, 190, 230, 235
- Bacon Francis, 61
- Ballintine, 61
- Bartlett E, 9
- Bearman JE, 63, 136, 139, 158–159
- Beck DF, 45
- Beecher HK, 21, 23, 57, 73
- Belson W, 200
- Benedict GW, 40–41
- Bennett AE, 148, 198, 200
- Bergman AB, 171
- Bernard C, 19
- beta-adrenoceptor blocking drugs, 228–230
- beta blocker heart attack trial, 230
- bias, due to
- analysis, 64
 - faulty methods, 66
 - inadequate control group, 56–59
 - noncompliance, 59–61
 - observer, 61–64
 - patient, 61
 - placebo effect (*see* placebo effect)
 - regression to the mean, 58
 - time trends, 58
 - uneven allocation, 56
 - withdrawal, 64–66
- biological importance, 29–30
- Black D, 243
- Blondlot, 61–62
- blood pressure measurement, 62–63, 82, 87–90, 94–95
- Böttinger LE, 215
- Bradford Hill A, vii, 1, 6, 11, 15–16, 22, 56, 74, 76–77, 191
- Braunwald E, 224
- Breslow NE, 188, 190
- Brown BW, 112
- Bull JP, 5–6, 9–10
- Bulpitt CJ, 30, 57, 85, 126, 146, 166, 170, 180, 198–199, 202, 205–206, 215, 236, 243
- Bush JW, 206
- Byar DP, 59, 134–135
- Cannel CF, 201
- carcinoma, bronchus, 2
- Carne S, 85, 146, 198–199, 202, 205
- Caron HS, 171
- carry-over effect, 119, 121, 126–128
- case-saving rule, 115
- Chalmers J, 126
- Chalmers TC, 134
- checking
- diagnoses, 91
 - documents, 90–93
 - measurements, 91–92, 186
 - outliers, 186
 - range, 93, 186
- chi-squared test, 187–189
- cholesterol measurements, 81
- Christie D, 59, 240
- Claimed Investigational Exemption, 209
- Clark CJ, 104, 115–116
- Clark HH, 151
- Clay PM, 241
- Clifton P, 170
- clinical trial certificate, 209
- clofibrate, 31–32, 217–218, 240
- Cobb LA, 23
- Cochrane AL, 2, 3, 243
- Cochran WG, 115, 127–128, 133, 179, 183–186
- coding of data, 152–156
- coefficient of variation, 81
- Collen MF, 199
- conduct of the trial
- nonadherence of investigators, 173–174
 - noncompliance of subjects, 170–173
 - pilot trial, 157–159
 - run-in period, 159–164
 - terminating, 175–178
 - withdrawals, 165–170
- confidence limits
- calculation, 189–191
 - rather than power, 104
- confounding variables, 46–47
- conjugated oestrogens, 26
- Conolly ME, 196
- consent
- avoidance of necessity for, 20
 - full information, 17–20, 75–76
 - written, 17, 20
- control group
- concurrent, 2, 7, 9–10
 - historical, 10–11, 58–59
 - similar to treatment group, 44–55
- Cooper GR, 81
- Co-operative Randomised Controlled Trial, 166–170
- coordinating centre, 137
- Cornfield J, 232
- Coronary Drug Project Trial, 26, 104, 114–115, 217–218, 227
- Cox DR, 190
- Cranberg L, 240
- Crockett JE, 23
- Croke G, 42
- Cronbach LJ, 201
- cross-over trial, 44

- effect of order, 121–123
- effect of treatment-period interaction, 121
- in psychiatry, 204
- reduction in numbers required, 113
- reduction in variability, 94
- to detect an interaction, 125–126
- when to use, 121–123
- Crout JR, 212
- Crowley FA, 90
- Cruickshank JM, 236
- data transformation, 186
- Day JL, 201
- decision-making trial, 108–109
- decision rules for stopping a trial, 25–26
- Declaration of Helsinki, 14
- definition of a randomised trial, vii
- designs for trials
- cross-over, 121–123
 - factorial, 123–126
 - graco-latin square, 128
 - latin square, 126–128
 - play-the-winner, 134–135
 - sequential, 128–134
 - standard 119–120
- design of documents, 148–150
- duplication, 150
 - instructions to subjects, 149–150
 - layout of forms, 148–149
 - no-carbon-required paper, 150
- Desu MM, 108
- detection of an ADR
- according to frequency, 214–216
 - in a large trial, 217–219
 - in a small trial, 217
 - methods, 219–220
 - when previously recognised, 214
- Devloo RA, 57
- dextrothyroxine, 26
- diabetes mellitus, 3, 26, 230–233
- dietary advice, 66–69, 172–173
- different trial designs, 118–135
- digit preference, 89
- Dimond EG, 23
- diphtheria, 10
- Doll R, 3, 59–60, 215, 219, 241
- Dollery CT, 58, 77, 85, 93, 146, 198–199, 202, 205, 210
- Dore CF, 89
- double-blinding technique, 69–71
- breakdown in, 72–74
 - definition, 61
 - disadvantages, 71–72
 - laboratory results, 70
 - method, 67–68
 - same treatment, 70
- when difficult, 68–69
- when undesirable, 68
- double-dummy technique, 71–72
- Douglas-Jones AP, 236
- Downie CC, 104, 115–116
- dropout (*see* withdrawal)
- drug interactions, 17
- Duncan JA, 200
- Dunnett CW, 115
- early trials on new drugs, 17–19, 209–213
- phase I, II and III trials, 210–211
- Ederer F, 57, 61, 64, 87, 190
- Elwood PC, 31–32, 226–227
- end point
- continuously distributed variable, 30
 - ease of measurement, 29
 - frequent, 31
 - identification, 28–29
 - infrequent, 31
 - qualitative, 30–31
- error
- alpha and beta, 99–112
 - analytical, 180–185
 - clerical, 93
 - diagnostic, 91
 - gamma, 107
 - in entry criteria, 91
 - in withdrawal of patients, 92–93
 - measurement, 91–92
 - one-tailed test, 183
 - types I and II, 99–112
 - type III, 107–109
- Etang JC L', 5
- ethics, 12–27
- declarations, 13–17
 - definition, 12
 - legal considerations, 13
 - nontherapeutic research, 14–15, 17
 - research combined with care, 14
 - research ethical committees, 16–17
 - sequential trials, 131
- encephalins, 58
- European Working Party on Hypertension in the Elderly, 23–25, 34, 161–162, 166, 176
- Evans W, 57
- Exclusion criteria, 23–24
- explanatory trial, 61
- Eysenck HJ, 204
- factorial trials, 123–126
- failure to accept results
- adverse effects, 237
 - faulty analysis, 223, 234–235
 - faulty definitions, 222–223, 235
 - faulty interpretation, 236

failure to accept results—*Continued*

- faulty performance, 230, 232–234
- faulty reporting, 223
- preconceived ideas, 226, 230–234
- restricted selection, 225, 235
- results not consistent
 - in subgroups 236
 - in different trials, 230, 237
- too few patients, 230, 235
- treatment groups different, 226, 232, 235
- treatment too difficult, 226, 228, 237
- treatment effect too small; 227–228, 235
- vested interests, 224, 237

- Fanshel S, 206
- false negative rate, 92, 199
- false positive rate, 92, 199
- Feinstein AR, 170, 172–174, 181, 230–234
- Ferris FL, 87
- Fibiger J, 10
- financial considerations, 139
- Fincke, M, 242
- Fish RC, 23
- Fisher, RA, 5, 123, 125
- Flamant R, 97, 106, 109–110
- Fletcher CM, 63
- Foulds GA, 63–64
- Free SM, 121
- Freiman JA, 179, 184

- Gail M, 115
- Gallstone Study (National Co-operative), 42–43
- Galton F, 86
- general practice, trials in, 4
- George SL, 108
- Gibb WE, 196
- Glantz SA, 179, 181, 184
- Glaser EM, 12, 18–19, 147, 159
- Goldberg ID, 85
- Goldsmith CH, 172
- Goldzieher JW, 242
- Gordis L, 170
- Gore SM, 179, 191
- Gorham GR, 204
- Graeco-Latin square, 128–129, 169
- Gravenstein JS, 57
- Green Governor DH, 18–19
- Greenwood DT, 209, 211
- Greiner T, 57
- Griffen JP, 210
- Grymes TP, 23
- Gullen WH, 63, 158–159

- Haenszel W, 241
- Halperin M, 107
- Hamilton M, 102–104, 144, 203–204
- Hamilton PJS, 35, 59–60

- Hampton JR, 162–163, 230
- Hargreaves MA, 241
- Harris AD, 125
- Harris WH, 236
- Havard CWH, 70, 126
- Hawthorne effect, 57
- Haybittle J, 144, 150–152
- Haygarth J, 8
- Healy MJR, 186
- Heaton-Ward effect, 73
- Hefferman A, 196
- Herxheimer A, 179, 213
- Higgin's law, 87
- Hill C, 135
- Hillis BR, 57
- Hills M, 120
- history of trials, 5–11
- Hoffbrand BI, 170, 205
- Hogben L, 180
- Holland WW, 90
- Hoyle C, 57
- Hsi BP, 171
- Hunter J, 7
- Hypertension Detection and Follow-up Program, 37, 39, 117, 192, 236
- hypoglycaemic drugs, 2, 26–27

- inclusion criteria, 23
- information to be collected, 144–156
 - demographic, 146
 - facilitating data entry, 145
 - identifying the patient, 145–146
 - primary care physician, 147
 - quantity, 144–145
 - to facilitate follow-up, 146
- intention-to-treat principle, 20, 46, 61, 173, 191–192, 234
- interaction between treatments, 123–126
- interim analyses, 25–26
- internal mammary artery ligation, 23
- International Anticoagulant Review Group, 32, 78, 228
- International Pharmaceuticals Substances Classification, 151

- Jellinek EM, 57
- Jenner, E, 7
- Jesty B, 7
- John JA, 186
- Jones G, 209–210
- Jones IG, 179

- Kahn HA, 63, 91
- Kahn RL, 201
- Karten I, 179, 181
- Keats AS, 57
- Kellner R, 204

- Kittle CF, 23
- Kolata GB, 224–225
- Knatterud GL, 217
- Knowleden J, 67
- Laplace PS, 9
- Lasagne L, 57
- latin square
 - carry-over effects, 217
 - order effects, 127
 - randomisation in, 126–127
 - treatment effect, 127
 - withdrawals, 169
- Lellouch J, 99, 106, 108–110
- Lewis JA, 104, 190, 216, 230, 235
- life table, 187–189
- Lilienfeld AM, 170
- Lind J, 6, 45
- Lionel NDW, 179
- Lionel W, 213
- Lipman RS, 170
- Lister J, 9–10
- Loewenson RB, 63, 158–159
- London Hospital Clinical Trials Unit, 140
- Long JR, 210
- Louis PCA, 9
- Lovell MG, 23
- LRC Coronary Prevention Trial, 40–42
- Lumbroso A, 211–212
- Lydecker M, 212
- Lynch P, 126

- Maccoby EE, 200
- Maccoby N, 200
- McFate Smith W, 136–137
- McMahon BT, 10
- McMahon FG, 196
- McMichael, 191
- McNemar's test, 187–188
- McPherson K, 26, 111–112
- Maitland C, 7
- Mann RD, 217
- Mantel N, 188, 241
- marketing trials, 4
- Markowitz M, 170
- Marple CD, 45
- masking, 61
- Maxwell C, 241
- Medical Research Council, 3, 5, 10–11, 16, 22–23, 73, 77, 162, 171, 203
- Meier P, 39, 121, 134, 177–178, 224
- Mellner C, 200
- methods of recruitment
 - medical records, 40–41
 - own patients, 40
 - screening, 41–42

- via colleagues, 40
- methyl dopa, 4, 195–196
- minor objectives
 - in long-term trials, 33–34
 - limiting factors, 34
- Mitchell GAG, 10
- Mitchell JRA, 225, 228
- Mittra B, 241
- monitor, 93–94, 174
- Mosteller FC, 57
- Muench's laws, 43, 63, 158–159
- Murphy's law, 157
- Mushlin AI, 171
- myocardial infarction, 2, 113–114, 176–177

- Nam JM, 115
- National Diet Heart Study, 68, 73, 104, 108, 113–115
- Nebuchadnezzar's trial, 5
- new treatment, 2, 17–19
- Nie NN, 186
- Nicholls DP, 146, 217
- nonadherence to protocol by investigator, 174–175
- noncompliance
 - assessment by interview, 170
 - assessment by pill count, 171
 - assessment by physician, 171
 - assessment by self-administered questionnaire, 170
 - assessment from blood and urine, 171
 - consequences, 172–173
 - detection, 162, 170–171
 - producing bias, 59–61, 172–173
- Norwegian Multicentre Study Group, 163–164, 177
- number of subjects required, 96–117
 - aids to calculating, 115–116
 - explanatory trial, 106–108
 - pragmatic trial, 108–110
 - relationship to asymmetric α , 110–111
 - drop out, 114
 - error, 109–112
 - power, 109–110
 - treatment effect, 113–114
 - variance, 112–113, 116–117
 - unequal groups, 114–115
 - with noncompliance, 172
- Nuremberg Code, 13

- Oates JA, 196
- objectives of a trial, 28–34
 - identification, 28–29, 97–99
 - importance, 29
 - likelihood to be achieved, 29
 - major, 28–33, 96–99

- objectives of a trial—*Continued*
 - quantification, 30–31
 - traditional, 97
- observer bias, 89
- observer training, 88–89
- Oliver MF, 217–219
- order effect, 122–123
- organization of a trial, 137
- Overall JE, 204
- Paré A, 5
- Park LC, 170
- Pasteur, 10
- Parten MB, 202
- patient as his own control, 44
- Pearson RM, 70, 126
- penicillin, 10, 21
- Perkin's tractors, 8
- Persantine-Aspirin Reinfarction Study, 227
- Peto R, 3, 43, 53, 58–60, 114–115, 179–180, 185, 187–188, 193, 241
- Petrie A, 179, 183, 185–186
- phase I, II & III trials, 209–210
- phenformin, 26
- physiotherapy, 2
- pilot trial,
 - advantages, 158–159
 - disadvantages, 159
- Pinner M, 10
- placebo effect, 22, 57–58
 - in hypertensive patients, 74–75
- placebo run-in period, 76
- placebos
 - advantages of, 74–75
 - disadvantages of, 77–78
 - doctor-patient relationship, 77
 - ethical considerations, 20–22, 77–78
 - in hypertension, 22, 65
 - in nineteenth century, 8
 - in pain relief, 21, 57–58
 - medico-legal problems, 77–78
 - number of trials required, 78
 - use of, 19, 64–65, 74–78
- play-the-winner trial, 54, 134–135
- Pocock SJ, 10, 240
- Pollock MR, 125
- Porrit AE, 10
- postmarketing trials, 212–213
- postoperative pain, 21, 57
- power (see error, type II)
- Pozdena RF, 62
- precision of a measurement, 81–83
- Prichard BNC, 196
- primary prevention trial, 31
- protocol
 - adherence to, 89–90
 - checklist, 10, 112
 - deviations from, 87–88
 - writing, 136–143
- Prout TE, 43
- Pryce IG, 242
- psychiatry
 - diagnosis, 203
 - informed consent, 203
 - measurement of response, 203
 - self-rating scales, 203
 - new drugs, 203
- psychotherapy, 2
- quality of life
 - factors influencing, 205
 - health status index, 206
 - measurement, 205–207
- questionnaire
 - interviewer administered, 201–202
 - self-administered,
 - advantages, 202
 - back translation, 202
 - validity, 92
- questionnaire design
 - characteristics of a good question, 150–151
 - characteristics of a good response, 151–152
 - coding response, 152–154
 - order of yes/no response, 152
 - responses in boxes, 152
- questions
 - clear, 198
 - grammatically correct, 200
 - high response rate, 200
 - only one enquiry, 200
 - open or closed 200–201
 - repeatable, 198
 - valid, 198–199
- quality control,
 - a clear protocol, 87–88
 - and double-blindness, 174
 - checking of documents, 90–93
 - entry criteria, 91, 174
 - training of observers, 88–90
 - trial monitor, 93–94, 174
 - use of different observers, 90
- random number tables, 49–50
- randomisation, 2, 10–11, 44–45
 - and acceptance of trial results, 47
 - and valid statistical tests, 47
 - advantages of, 45–47
 - alternatives to, 54
 - balanced block, 50
 - block, 49
 - definition, 44
 - disadvantages of, 47–48
 - ethical problems with, 48

- restricted, 47, 49–51
- stratified, 48, 53
- tests for, 185
- timing of, 49
- unequal, 53–54, 114–115
- variable block, 51–52
- recruitment of subjects, 39–43
 - methods (see methods of recruitment)
 - period of recruitment, 40
 - related factors, 42–43
 - selection criteria, 39–40
- reduction in variance
 - by measuring change, 112–113
 - by replication, 112–113
 - in cross-over trial, 113
- regression to the mean, 58, 86
- regulatory approval
 - for early drug trials, 209–210
 - for general release, 211–212
- repeatability indices, 83–86
- repeatability of a measurement, 80–87
 - and bias, 86
 - and regression to the mean, 86–87
 - and validity, 87
 - qualitative data, 83–86
 - quantitative data, 80–83
 - symptom reports, 85
- repeated looks, 25–26, 111–112, 234
- SPSS, 186
- square-root rule, 115
- Stamler J, 39
- standard deviation and standard error, 184–185
- steering committee, 137
- streptokinase, 225–226
- streptomycin, 3, 10, 16, 23
- subjective well-being, 194–208
- sulphinpyrazone, 31–32
- Sunday Times, 4, 212–213
- Sutton HG, 8
- Sweetnam PM, 227
- symptoms
 - recognition, 194–197
 - spontaneous reporting, 197
- synergism, 125
- Taylor JA, 204
- Temple R, 224–225
- terminating a trial
 - decision rules, 25–26, 175–178
 - due to an adverse effect, 26
 - due to benefit, 26
 - stopping too late, 177
 - stopping too soon, 176
- Thompson FM, 102–104

- Todd AH, 209, 211
- tolbutamide, 26, 230–233
- tonsillectomy, 3
- toxicity, detection of, 26
- Travell J, 57
- treble blindness, 61
- Twenty-one Army Group, 10
- two-armed-bandit problem, 135
- Uberla K, 227
- unequal groups (see randomisation)
- University Group Diabetes Project, 26–27, 230–233
- vaccination against smallpox, 7–8
- validity
 - and repeatability, 87
 - definition, 35
 - extrapolation of results, 36–38
 - of questions, 198–199
- variability of results, 80–95, 105–106 (see accuracy, precision and repeatability)
- variance
 - qualitative data, 105–106
 - quantitative data, 105
- Vedin JA, 228–230
- Vere DW, 217
- Veterans Administration trials, 22, 26, 30–31, 36–37, 165, 205
- Vitamin C, 73
- volunteers,
 - prisoners, 18–19
 - recruitment, 18
 - remuneration, 17–18
- Wade OL, 13
- Wald A, 129
- wash-out period, 121
- Weinstein MC, 47–48, 53
- Werner RJ, 171
- Westerholm B, 215
- Whitehead T, 30, 186
- Whitlock RM, 229
- Wilhelmsson C, 229
- Wilkinson GN, 170
- Williams EF, 127
- Williams, OD, 93
- Williams W, 7
- Wilson C, 125
- Wilson EB, 63
- Wisniewski TK, 102–104
- withdrawal from a trial, 24–25, 138, 165–169
 - due to adverse treatment effect, 166
 - due to death, 165
 - due to worsening of condition, 166–167
 - in activity treated group, 65–66
 - in error, 92, 93

withdrawal from a trial—*Continued*
in run-in period, 64–65, 162, 167
methods of analysing, 165, 168–169
methods of reducing number, 167–169
related to end point, 165–166
unrelated to end point, 165

Wood, Doctor, 62

Woolf B, 241

World Medical Association, 14–15

Wortley, Lady M, 7

Wright BM, 89

Wright IS, 45

Wright P, 144, 150–152

Wulff HR, 190

Young DW, 199

Zar JH, 182

Zelen M, 20, 44, 47, 50, 134