2.3.3 Large-scale randomized evidence: trials and meta-analyses of trials

C. Baigent, R. Peto, R. Gray, S. Parish, and R. Collins

Essentials

Reliable detection or refutation of realistically moderate effects on major outcomes often requires large-scale randomized evidence

As long as doctors start with a healthy scepticism about the many apparently striking claims and counter-claims that appear in the medical literature, trial results do make sense. The main enemy of common sense is over-optimism: there are a few striking exceptions where treatments for serious disease work extremely well, but many claims of vast improvements from new therapies turn out to be evanescent.

Clinical trials generally need to be able to detect or to refute realistically moderate (but still worthwhile) differences between treatments in long-term disease outcome. Large-scale randomized evidence should be able to detect such effects, but medium-sized trials or medium-sized meta-analyses can, and often do, yield false negative or exaggeratedly positive results. If the results of such studies seem too good to be true then they probably are conversely, unpromising evidence can be misleading if it is from a study of inadequate size, or from one particular subgroup of a large study with a clearly favourable overall result. Realistically moderate expectations of what a treatment might achieve (or, if one treatment is to be compared with another, of how large any difference between the main effects of these two treatments is likely to be) should foster studies that can discriminate reliably between (1) a difference in outcome that is realistically moderate but still worthwhile, and (2) a difference in outcome that is too small to be of any material importance.

To assess moderate effects reliably, avoid both moderate biases and moderate random errors

To demonstrate or refute realistically moderate differences in outcome, studies must guarantee both (1) strict control of bias - which, in general, requires proper randomization and appropriate statistical analysis, with no unduly 'data-dependent' emphasis on
specific parts of the overall evidence; and (2) strict control of the play of chance - which, in general, requires large numbers with the outcome of interest, rather than a lot of detail on each patient. The conclusion is obvious: moderate biases and moderate random errors must both be avoided if moderate benefits are to be assessed reliably. This leads to the need for large numbers of properly randomised patients with properly analysed data, which in turn should lead to some large but simple randomized trials (‘mega-trials’) and to large systematic overviews (‘meta-analyses’) of all related randomized trials.

Other forms of evidence may be untrustworthy

Non-randomized evidence, unduly small randomized trials, unduly small meta-analyses of trials and undue emphasis on particular subgroups (or on particular trials) are all much inferior as sources of evidence about current patient management or as foundations for future research strategies because they often cannot discriminate reliably between moderate (but worthwhile) differences and negligible differences in outcome, and the mistaken clinical conclusions that they engender could well result in the undertreatment, overtreatment, or other mismanagement of millions of future patients worldwide.

Benefits of large-scale randomized evidence

In contrast, many premature deaths each year could be avoided by seeking appropriately large-scale randomized evidence about various widely practicable treatments for the common causes of death, and by disseminating this evidence appropriately. The value of such large-scale randomized evidence is illustrated by the trials of fibrinolytic therapy for acute myocardial infarction; of anti-platelet therapy for a wide range of vascular conditions; of hormonal therapy for early breast cancer; and of drug therapy for lowering blood pressure. In these examples, proof of benefit that could not have been achieved by either small-scale randomized evidence or non-randomized evidence of benefit has led to widespread changes in practice that are now preventing hundreds of thousands of premature deaths each year, and appropriately large-scale randomized evidence could substantially improve the management of many important, but non-fatal, medical conditions.

Moderate (but worthwhile) effects on major outcomes are generally more plausible than large effects

Some treatments have large, and hence obvious, effects on survival; e.g. it was clear without the need for any randomized trials that prompt treatment of diabetic coma or cardiac arrest can save lives, and more recently the introduction of protease inhibitors for the treatment of HIV infection led to a reduction in AIDS-related morbidity and mortality that was large enough to be obvious even without randomized evidence; indeed, the remarkable effectiveness of antiretroviral drugs can be seen from the sudden reversal, after the mid 1990s, of the upward trend in mortality among men aged 30–34 in the United States of America (Fig. 2.3.3.1), the chief cause of which was HIV/AIDS.

However, over the past few decades the hopes of large treatment effects on mortality and major morbidity in many serious diseases have been unrealistically high. Of course, treatments do quite commonly have large effects on various less fundamental measures: certain drugs clearly reduce blood pressure, blood cholesterol, or blood glucose; many tumours or leukaemias in middle and old age can be controlled temporarily by radiotherapy or chemotherapy; and, in acute myocardial infarction, lidocaine (lignocaine) can prevent many arrhythmias and fibrinolytic therapy can dissolve many thrombi. However, although such effects on intermediate outcomes may be large, the net effects on mortality may be much more modest.

In general, if substantial uncertainty remains about the efficacy of a practicable treatment, its effects on major endpoints are probably either negligibly small, or only moderate, rather than large. Indirect support for this rather pessimistic conclusion comes from many sources, including: the previous few decades of disappointingly slow progress in the curative treatment of common chronic diseases of middle age; the heterogeneity of each single disease, as evidenced by the unpredictability of survival duration even when apparently similar patients are compared with each other; the variety of different mechanisms in certain diseases that can lead to death, only one of which may be appreciably influenced by any one particular therapy; the modest effects often suggested by meta-analyses (see later) of various therapies, and, in certain special cases, observational epidemiological studies of the strength of the

---

**UNITED STATES 1950–2004: Males & Females**

All medical mortality at ages 30–34

![Mortality trends in the United States of America among men and women aged 30–34 during the period 1950–2004. Antibacterial drugs caused a big decrease in mortality around the middle of the century in both sexes. The increase in AIDS-related mortality since the early 1980s caused a sharp increase in all-cause mortality, particularly in men, which continued until it was spectacularly reversed by effective antiretroviral drug combinations in the mid 1990s.](Fig. 2.3.3.1)
relationship between some disease and the factor that treatment will modify (e.g. blood pressure, blood cholesterol, or blood glucose: see later).

Having accepted that only moderate reductions in mortality are likely with many currently unevaulated interventions, how worthwhile might such effects be if they could be detected reliably? To some clinicians, reducing the risk of early death in patients with myocardial infarction from 10 per 100 patients down to 9 or 8 per 100 patients treated may not seem particularly worthwhile, and if such a reduction was only transient, or involved an extremely expensive or toxic treatment, this might well be an appropriate view. Worldwide, however, several million patients a year suffer an acute myocardial infarction, and if just one million were to be given a simple, nontoxic, and widely practicable treatment that reduced the risk of early death from 10% down to 9% or 8% (that is, a proportional reduction of 10 or 20%), this would avoid 10000 or 20000 deaths. (At least 1 million patients a year now receive fibrinolytic therapy for acute myocardial infarction, which is avoiding about 20000 early deaths a year.) Such absolute gains are substantial, and might considerably exceed the number of lives that could be saved by a much more effective treatment of a much less common disease.

**Reliable detection or refutation of moderate differences requires avoidance of both moderate biases and moderate random errors**

If realistically moderate differences in outcome are to be reliably detected or reliably refuted, then errors in comparative assessments of the effects of treatment need to be much smaller than the difference between a moderate, but worthwhile, effect and an effect that is too small to be of any material importance. This in turn implies that moderate biases and moderate random errors cannot be tolerated. The only way to guarantee very small random errors is to study really large numbers, and this can be achieved in two main ways: by making individual studies large, and by combining information from as many relevant studies as possible in a systematic meta-analysis (Box 2.3.3.1). However, it is not much use having very small random errors if there may well be moderate biases, so even the large sizes of some nonrandomized analyses of computerized hospital records, where the complex factors involved in the decision to treat a person with a particular drug may not be recorded in sufficient detail, cannot guarantee medically reliable comparisons between the effects of different treatments (see later). For, the choice of treatment may be strongly affected by subtle patient characteristics that are correlated with the prognosis. (A crude illustration of such problems is provided by the old joke ‘What’s the most dangerous place in the world? ’Bed—look at the number of people who die in bed!’.)

**Avoiding moderate biases**

Proper randomization avoids systematic differences between the types of patient in different treatment groups. The fundamental reason for randomization is to avoid moderate bias, by ensuring that each type of patient can be expected (but for the play of chance) to have been allocated in similar proportions to the different treatment strategies that are to be compared. This means that only random differences should affect the final comparisons of outcome. Nonrandomized methods, in contrast, cannot generally guarantee that the types of patient given the new study treatment do not differ systematically in any important ways from the types of patient given any other treatment(s) with which the new study treatment is to be compared. For example, moderate biases may arise if the study treatment is novel and doctors are afraid to use it for the most seriously ill patients, or, conversely, if they are more ready to use it for those who are desperately ill. There may also be other ways in which the severity of the condition differentially affects the likelihood of being assigned to different treatments by the doctor’s choice (or by the patient’s choice, or by any other nonrandom procedure).

It might appear at first sight that by collecting enough information about various prognostic features it would be possible to make some mathematical adjustments that correct for any such differences between the types of patients who, in a nonrandomized study, receive the different treatments that are to be compared. The ill-conceived hope is that such methods, which are often carried out on routinely collected health care data, might achieve comparability between those entering the different treatment groups, but they cannot be guaranteed to do so, and often fail seriously. The difficulty is that some important prognostic factors may be unrecorded, while others may be difficult to assess exactly and hence difficult to adjust for reliably. Although there are examples of nonrandomized studies in which the estimated effects of treatment appear quantitatively close to those observed in analogous randomized trials, there are many examples where they do not, being either quantitatively incorrect—so that drugs appear either misleadingly promising or of misleadingly low efficacy—or even qualitatively incorrect, when a harmful drug might appear effective (or vice versa).

---

**Box 2.3.3.1 Requirements for reliable assessment of moderate effects: negligible biases and small random errors**

**Negligible biases**

(i.e. guaranteed avoidance of moderate biases)

**Proper randomization**

(nonrandomized methods might suffer moderate biases)

**Analysis by allocated treatment**

(including all randomized patients: ‘intention to treat’ analysis)

**Chief emphasis on overall results**

(no unduly data-dependent emphasis on particular subgroups)

**Systematic overview of all relevant randomized trials**

(no unduly data-dependent emphasis on particular studies)

**Small random errors**

(i.e. guaranteed avoidance of moderate chance fluctuations)

**Large numbers in any new trials**

(to be really large, trials should be ‘streamlined’)

**Systematic overview of all relevant randomized trials**

(whcih yields the largest possible total numbers)
The machinery of a properly randomized trial

No foreknowledge of what the next treatment will be

In a properly randomized trial, the decision to enter a patient is made in ignorance of which of the trial treatments that patient will, once entered, be allocated. The treatment allocation is then made known after trial entry has been decided upon. (The purpose of this sequence is to ensure that foreknowledge of what the next treatment is going to be cannot affect the decision as to whether to enter the patient; if it did, those to be allocated one treatment might differ systematically from those to be allocated another.) Ideally, any major prognostic features should also be irreversibly recorded before the treatment is revealed, particularly if these are to be used in any treatment analyses. For, if the recorded value of some prognostic factor might be affected by knowledge of the trial treatment allocation, then treatment comparisons within subgroups defined by that factor might be moderately biased. In particular, treatment comparisons just among ‘responders’ or just among ‘nonresponders’ can be misleading unless the response is assessed before treatment allocation (which it can sometimes be, if all patients have a ‘run-in’ period on active treatment before randomization, partly to assess the response to treatment and partly to exclude those who seem during this prerandomization run-in unlikely to participate wholeheartedly in the main post-randomization study).

No bias in patient management or in outcome assessment

An additional difficulty, in both randomized and non-randomized comparisons of various treatments, is that there might be systematic differences in the use of other treatments (including general supportive care) or in the assessment of major outcomes. A non-randomized comparison may well suffer from moderate biases due to such systematic differences in ancillary care or assessment, particularly if it merely involves the retrospective review of medical records. In the context of a randomized comparison, however, it is generally possible to devise ways to keep any such biases small. For example, placebo tablets may be given to control–allocated patients and certain subjective assessments may be ‘blinded’ (although this may be less important in studies assessing mortality).

‘Intention-to-treat’ analyses with no post-randomization exclusions

Even in a properly randomized trial, unnecessary biases may be introduced by inappropriate statistical analysis. One of the most important sources of bias in the analysis is undue concentration on just part of the evidence; that is to say, on ‘data-derived subgroup analyses’ (see below). Another easily avoided bias is caused by the post-randomization exclusion of patients, particularly if the type (and hence prognosis) of those excluded differs from one treatment group to another. Therefore one of the fundamental statistical analyses of a trial that should be made available is an analysis that compares all those originally allocated one treatment (even though some of them may not have actually received it) with all those allocated the other treatment. This is sometimes referred to as an ‘intention-to-treat’ analysis. Additional analyses can also be reported: e.g. in describing the frequency of some very specific side effect it may well be preferable to record its incidence only among those who actually received the treatment. (This is because strictly randomized comparisons may not be needed to assess extreme relative risks.) However, in assessing moderate effects on the main outcome of interest such ‘on-treatment’ analyses can be misleading, and ‘intention-to-treat’ analyses are generally a more trustworthy guide as to whether there is any real difference between the trial treatments in their effects on long-term outcome.

Unduly data-dependent emphasis on results in particular subgroups

Treatment that is appropriate for one patient may be inappropriate for another. Ideally, therefore, what is wanted is not only an answer to the question, ‘Is this treatment helpful on average for a wide range of patients?’, but also an answer to the question, ‘For which recognizable categories of patient is this treatment particularly helpful?’ However, this ideal is difficult to attain directly because apparently differences between the proportional risk reductions in particular subgroups of patients are often surprisingly unreliable. Of course, patients who already have a very good prognosis anyway and are at low absolute risk cannot have a large absolute benefit (for even if a small risk is halved the absolute benefit is small).

Classification of patients as being at low (or high) risk of an adverse disease outcome is often a useful guide as to which patients can expect little absolute gain even if the trial treatment works as expected (and as to which patients might expect a worthwhile gain). This low risk/high risk split may not require support from formal subgroup analyses—indeed, it could even be damaged by such analyses. For, even if the proportional effects of treatment in specific subgroups are importantly different, standard subgroup analyses are so insensitive that they may well fail to demonstrate these differences. Moreover, even if there are highly significant differences between the proportional risk reductions produced by the trial treatment in different subgroups, and the results seem to suggest that the treatment works in some subgroups but not in others (thereby giving the appearance of a ‘qualitative interaction’), this may still not be good evidence for subgroup–specific treatment preferences. The play of chance often produces qualitatively wrong answers in particular subgroups in trials (or in meta-analyses of trials) that could, if interpreted incautiously, lead to millions of people being treated inappropriately, or untreated inappropriately.

Questions about such ‘interactions’ between patient characteristics and the effects of treatment are easy to ask, but are surprisingly difficult to answer reliably. Apparent interactions can often be produced by the play of chance and, in particular subgroups, can mimic or obscure some of the moderate treatment effects that might realistically be expected. To illustrate this, a subgroup analysis was performed based on the astrological birth signs of patients randomized in the very large Second International Study of Infarct Survival (ISIS-2) trial of aspirin for suspected acute myocardial infarction. Overall in this trial, the 1-month survival advantage produced by aspirin was conclusively demonstrated (804 vascular deaths among 8587 patients allocated aspirin vs 1016 among 8600 allocated no aspirin; 23% proportional reduction, two-sided p value <0.000001). However, when these analyses were subdivided into 12 subgroups by the patients’ birth signs (in medieval Western astrology, the ‘birth sign’ is determined by the month of birth: e.g. ‘Libra’ means born 24 September to 23 October, and ‘Gemini’ means born 22 May to 21 June) to illustrate...
the unreliability of subgroup analyses, aspirin appeared totally ineffective for those born under Libra or Gemini (Table 2.3.3.1). If it would obviously be unwise to conclude from such a result that patients born under the astrological birth sign of Libra or Gemini should not be given aspirin if they have a heart attack. However, similar conclusions based on ‘exploratory’ data-derived subgroup analyses, which, from a purely statistical viewpoint, are no more reliable than these, are often reported and believed, with inappropriate effects on worldwide clinical practice.

There are three main remedies for this unavoidable conflict between the reliable subgroup-specific conclusions that doctors and patients want and need, and the statistically unreliable findings that direct subgroup analyses can usually offer. However, the extent to which these remedies are helpful in particular instances is one on which informed judgements differ.

First, where there are good a priori reasons for anticipating that the proportional effects of treatment might be very different in different circumstances then a limited number of subgroup analyses may be performed in the study protocol, along with a prediction of the direction of such proposed interactions. (For example, it was expected that the benefits of fibrinolytic therapy for acute myocardial infarction would be greater the earlier such patients were treated and so some studies prespecified that the analyses would be taken somewhat more seriously than other subgroup treatments of heart attack in the 1980s, and many other mega-trials have now been successfully undertaken, not only in the field of cardiology—where numerous large trials have already been performed—but also in other specialties where treatment might be expected to have only moderate effects on morbidity and mortality from a common disease or injury. Many such mega-trials have produced medically important results that would not otherwise

### Table 2.3.3.1 False-negative mortality effect in a subgroup defined only by the medieval astrological birth sign: the ISIS-2 (1988) trial of aspirin among over 17,000 patients with acute myocardial infarction

<table>
<thead>
<tr>
<th>Astrological birth sign</th>
<th>No. of 1-month deaths (aspirin vs placebo)</th>
<th>Statistical significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Libra or Gemini</td>
<td>150 vs 147</td>
<td>NS</td>
</tr>
<tr>
<td>All other signs</td>
<td>654 vs 869</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Any birth sign*</td>
<td>850 (9.4%) vs 1016 (11.8%)</td>
<td>p &lt; 0.0001</td>
</tr>
</tbody>
</table>

*Appropriate overall analysis for assessing the true effect in all subgroups. Medieval astrologist divides birth dates into 12 ‘birth signs’ (which depend only on the day and month of birth, not the year of birth). To demonstrate the potential unreliability of other subgroup analyses, the ISIS-2 patients were divided into 12 subgroups according to their astrological birth sign, and the apparent effects of aspirin were calculated separately in each of these 12 subgroups. Because of the play of chance the apparent effects differed from one subgroup to another, ranging from no apparent effect of aspirin in two subgroups (Libra and Gemini; see text for definition) to aspirin apparently halving mortality in another (Capricorn).

### Avoiding moderate random errors

#### The need for large-scale randomization

To distinguish reliably between the two alternatives that (1) there is no worthwhile difference in survival or that (2) treatment confers a moderate, but worthwhile, benefit (e.g. 10 or 20% fewer deaths), not only must systematic errors be guaranteed to be small (see above) compared with such a moderate risk reduction, but so too must any of the purely random errors that are produced just by chance. Random errors can be reliably avoided only by studying very large numbers of patients and hence large enough numbers of ‘endpoints’. However, it is not sufficiently widely appreciated just how large clinical trials need to be in order to detect moderate differences reliably. This can be illustrated by a hypothetical trial that is actually quite inadequate—even though by some standards it is moderately large—in which a 20% reduction in mortality (from 10% to 8%) is supposed to be detected among 2000 heart attack patients (1000 treated and 1000 controls). In this case, one might predict about 100 deaths (10%) in the control group and 80 deaths (8%) in the treated group. However, if this difference were to be observed it would not be conventionally significant (p = 0.1); indicating that even if there is no real difference between the effects of the trial treatments, it would still be relatively easy for a result at least as extreme as this to arise by chance alone. Although the play of chance might well increase the difference enough to make it conventionally significant (e.g. 110 deaths vs 70 deaths, 2p < 0.001), it might equally well dilute, obliterate (e.g. 90 deaths vs 90 deaths), or even reverse it. The situation in real life is often even worse, as the average trial size may include only a few dozen events rather than the several hundred (or few thousand) that would ideally be needed to guide the future treatment of millions.

#### Mega-trials: how to randomize large numbers

One of the chief techniques for obtaining appropriately large-scale randomized evidence is to make trials extremely simple, and then to invite hundreds of hospitals to collaborate. The first of these large streamlined trials (or mega-trials) were the ISIS and GISSI studies of heart attack treatment in the 1980s, and many other mega-trials have now been successfully undertaken, not only in the field of cardiology—where numerous large trials have already been performed—but also in other specialties where treatment might be expected to have only moderate effects on morbidity and mortality from a common disease or injury. Many such mega-trials have produced medically important results that would not otherwise
have been reliably obtained. However, in terms of medically significant findings, what has been achieved so far is only a fraction of what would be possible if this research strategy could be more widely adopted. Any obstacle to simplicity is an obstacle to large size, and so it is worth making enormous efforts at the design stage to simplify and streamline the process of entering, treating, and assessing patients. Many trials would be of much greater scientific value if they collected 10 times less information, both at entry and during follow-up, on 10 times more patients. Since those responsible for entering patients into trials are generally busy people, it is particularly important to simplify the entry of patients, otherwise rapid recruitment may prove difficult (see later). Likewise, when allocating resources within large-scale trials, it is important to direct them to where it chiefly matters, namely the recruitment of large numbers of patients and counting how many suffer the main outcomes of interest, instead of wasting large sums of money on inappropriate audits or unnecessary or excessively frequent measurements, the analysis of which will contribute little to answering the main study question.

Simplification of entry procedures for trials: the ‘uncertainty principle’

For ethical reasons, patients cannot have a commonly available treatment chosen at random for them if either they or their doctor are (for any reasons) already reasonably certain that another treatment is preferable. Hence, randomization can be offered only if both doctor and patient feel substantially uncertain as to which of the trial options is best. The question then arises, ‘Which categories of patients about whose treatment there is such uncertainty should be offered randomization?’ The obvious answer is all of them, welcoming the heterogeneity that this will produce. (For example, either the treatment of choice will turn out to be the same for men and women, in which case the trial might as well include both, or it will be different, in which case it is particularly important to study both.) In appropriately large trials, patient homogeneity is generally a defect while heterogeneity is generally a strength. Consider, for example, the trials of immediate fibrinolytic therapy for acute myocardial infarction. Some had restrictive entry criteria that allowed inclusion of only those patients who presented between 0 and 6 h after the onset of pain, and those trials contributed almost nothing to the key question of how late such treatment can still be useful. In contrast, the trials with wider and more heterogeneous entry criteria that included some patients with somewhat longer delays between pain onset and hospitalization were able to show that fibrinolytic therapy can have definite protective effects not only when given 0 to 6 h but also when given 7 to 12 h after the onset of pain (see later).

This approach of randomizing the full range of patients in whom there is substantial uncertainty as to which treatment option is best was used in the first Medical Research Council Asymptomatic Carotid Surgery Trial (ACST-1). Narrowing of the carotid artery (which is rapidly detectable by ultrasound) can eventually cause a stroke, or even a succession of strokes. It can be dealt with surgically by carotid endarterectomy, but in the 1990s there was much uncertainty as to whether such surgery, with its inherent perioperative risks, was appropriate for individuals with severe carotid artery narrowing who were currently asymptomatic (i.e. had not had a stroke in the past few months). The ACST was therefore designed to compare a policy of immediate carotid endarterectomy versus a policy of ‘watchful waiting’ in asymptomatic patients with substantial carotid artery narrowing. If a patient was prepared to at least consider surgery seriously, then the neurologist and surgeon responsible for that individual’s care considered in their own undefined way whatever medical, personal, or other factors seemed to them to be relevant, including, of course, the patient’s own preferences and values. Eligibility for randomization was defined by the ‘uncertainty principle’ (Fig. 2.3.3.2):

- If they or the patient were reasonably certain, for any reason, that they did wish to recommend immediate surgery for that particular patient, the patient was not eligible for entry into the ACST.
- Conversely, if they or the patient were reasonably certain, for any reason, that they did not wish to recommend immediate surgery, the patient was likewise not eligible for entry into the trial.
- If, but only if, the doctor(s) and patient were substantially uncertain what to recommend, the patient was automatically eligible for randomization between immediate vs no immediate surgery (with all patients receiving whatever their doctors judged to be the best available medical care, which generally included advice to stop smoking, low-dose aspirin, treatment of hypertension; and, in the latter years of the trial, a statin).

In ACST-1, there were substantial differences between individual doctors in the types of patient about whom they were uncertain (in terms of the severity of carotid stenosis [which was generally recorded on ultrasound as 70%, 80%, or 90% blockage], age, general health and various other characteristics). This guaranteed that no category of patient about which there was widespread uncertainty would be wholly excluded, and hence guaranteed that the trial would yield at least some direct evidence in a wide range of typical patients. As a result of the wide and simple entry criteria adopted by ACST-1, 3120 patients were randomized (which was more than in any previous vascular surgery trial), so the study was able to provide some clear answers about who needed carotid endarterectomy. In asymptomatic patients younger than 75 years of age, with carotid diameter about 70% or more on ultrasound, immediate carotid endarterectomy halved the net 5-year stroke risk from about 12% to 6% (even though this 6% included the 3% perioperative hazard). For patients with only moderate carotid artery stenosis on ultrasound, the 5-year risks of carotid stroke (excluding perioperative hazard) were 2% vs 9%, whilst among those with tighter stenosis the risks were 3% vs 10%, suggesting about as much benefit in moderate as in tight stenosis.

The ‘uncertainty principle’ simultaneously meets the requirements of ethicality, heterogeneity, simplicity and maximal trial size, and should be widely used. It states that the fundamental eligibility criterion is that both doctor and patient should be substantially uncertain about the appropriateness of each of the trial treatments for that particular patient. With such uncertainty as the fundamental criterion of eligibility, informed consent can often be simplified. For, the degree of ‘informed consent’ that is appropriate in a randomized comparison of two established treatments governed by the ‘uncertainty principle’ should probably not differ greatly from that which is applied in routine practice outside trials when treatment is being chosen haphazardly—or, to put it another way, ‘double standards’ between trial and non-trial situations are inappropriate. The haphazard nature of many
nonrandomized treatment choices is reflected in the wide variations in practice between and within countries. Even when a practice is similar, it may be similarly wrong: e.g. before the ISIS-2 results became available (see later), few doctors routinely used fibrinolytic therapy for acute myocardial infarction. Provided that trials are governed by the ‘uncertainty principle’, there is an approximate parallel between good science and good ethics. Indeed, in such circumstances, excessively detailed consent procedures (which can be distressing and inhumane, and so would not be considered appropriate in routine nontrial clinical practice) would not be humane or ethically appropriate in trials. Excessively detailed consent procedures are, unfortunately, quite common, but their chief purpose is to protect doctors against lawyers rather than to protect patients against anything.

This ‘uncertainty principle’ is just one of many ways of simplifying trials and thereby helping them to avoid becoming enmeshed in a mass of wholly unnecessary traditional complexity. If randomized trials can be substantially simplified (which, it must be admitted, requires a reversal of the current trend towards unnecessary complexity), as has already been achieved for a few major diseases, and hence made very much larger, then they will continue to play an appropriately central role in the development of rational criteria for planning treatment strategies and reducing death and disability.

Minimizing both bias and random error: meta-analyses of randomized trials

Archie Cochrane was one of the first people to emphasize the need to organize, by specialty, the results from all relevant randomized trials, and the Cochrane Library brings together in a single place a large number of systematic reviews (many of which include meta-analyses of randomized trials) summarizing the available evidence about a wide range of therapeutic questions. When several trials have all addressed much the same question, the traditional procedure of only a few of them becoming widely known may be a source of serious bias, since chance fluctuations for or against treatment may affect which trials become well known and widely cited.

Fig. 2.3.3.2 Example of the ‘uncertainty principle’ to define eligibility for trial entry: the chief eligibility criterion for the Asymptomatic Carotid Surgery Trial (ACST) was that doctors and patients should be substantially uncertain whether to risk immediate carotid surgery. Partly because this criterion was appropriately flexible, ACST-1 became the largest-ever trial of vascular surgery, showing that the long-term benefits of carotid artery surgery could eventually outweigh the immediate hazards. ACST-2 (http://www.acst.org) is now randomizing surgery vs carotid stenting where the doctor(s) and patient are substantially uncertain which to prefer.

Carotid artery stenosis detected by ultrasound, but, as yet, no clinical evidence of stroke from it.

Should patient be offered immediate cardiac surgery?

Doctor(s) or patient reasonably certain (no matter why) that immediate surgery is not appropriate: Ineligible

Doctor(s) or patient reasonably certain (no matter why) that immediate surgery is appropriate: Ineligible

Doctor(s) and patient substantially uncertain whether to risk immediate surgery. Uncertainty implies eligibility Telephone or fax to randomize

Group 1: Allocated No immediate surgery (unless or until a clear indication is thought to have arisen)

Group 2: Allocated Immediate carotid surgery (unless definite contraindication is thought to have been discovered, or patient/doctor changes their mind)

Over the next few years only a small number get carotid surgery

90% get carotid surgery (median delay: 1 month)

Statistical comparisons of various outcomes over 5 to 10 years 100% of Group 1 vs. 100% of Group 2, i.e. ‘intention to treat analysis’
To avoid this problem, it is appropriate to base inference chiefly on a meta-analysis of all the results from all of the trials that have addressed a particular type of question (or on an unbiased subset of such trials), and not on some potentially biased subset of these trials. Such meta-analyses will also minimize random errors in the assessment of treatment since, in general, far more patients are involved in a meta-analysis than in any contributory individual trial.

The separate trials may well be heterogeneous in their entry criteria, their treatment schedules, their follow-up procedures, their methods of treating relapse, etc. In view of this heterogeneity, at one extreme each trial might be considered in virtual isolation from all others, while at the opposite extreme the results from all trials could be combined, largely ignoring any heterogeneity. Both these extreme views have some merit, and the pursuit of each by different people may have more illuminating than too definite an insistence on any one particular approach. However, the heterogeneity of the different trials merely argues for careful interpretation of any meta-analyses of different trial results, rather than arguing against meta-analyses. For, whatever the difficulties in interpreting meta-analyses may be, without them it is difficult to avoid moderate selective biases and substantial random errors, both of which could obscure any moderate treatment effects, or, conversely, imply an effect where none exists.

**Which meta-analyses are trustworthy?**

Since the 1970s, a rapidly increasing number of meta-analyses of the results of randomized trials have been reported, not all of which are trustworthy. When considering how reliable a given one might be there are two fundamental questions: what is the potential for bias, and what is the potential size of purely random errors? To answer the first question consideration must be given to whether biases might exist within individual trials (e.g. because of an unreliable method of randomization or because of post-randomization exclusions from the main analyses), and whether the subset of trials under consideration might be a biased subset of all relevant trials that have been performed (as might arise, for example, if certain trials were abandoned because of unpromising findings, or remained unpublished for this or any other reason).

The simplest approach to meta-analysis is merely to have collected and tabulated the published data from whatever randomized trial reports are easily accessible in the literature, and sometimes this may suffice. At the opposite extreme, extensive efforts may have been made by those organizing the meta-analysis to locate every potentially relevant randomized trial, including those never published, to collaborate closely with the trialists to seek individual data on each patient ever randomized into those trials, and then (after extensive checks and corrections of such data) to produce, in collaborative re-analyses with those trialists, agreed analyses and publications. The results of some of the largest such collaborative re-analyses will be described later: the Anti-Thrombotic Trialists’ (ATT) Collaborative Group, the Fibrinolytic Therapy Trialists’ (FTT) Collaborative Group, and the Early Breast Cancer Trialists’ Collaborative Group (EBCTCG). Collaboration of the original trialists in the meta-analysis process, with collection of detailed data from each individual trial participant, can help to avoid or minimize the biases that could be produced by missing trials (e.g. owing to the greater likelihood of extremely good, or extremely bad, results being particularly widely known and published), by inappropriate post-randomization withdrawals, or by the failure to allocate treatment properly at random. If randomization was performed properly in the first place, the post-randomization withdrawals can often be followed up and restored to the study for an appropriate ‘intention-to-treat’ analysis. Knowledge of the exact methods of treatment allocation (backed up by checks on whether the main prognostic factors recorded are nonrandomly distributed between the treatment groups in a particular trial) may help to identify trials that were so improperly randomized that they should be excluded from a meta-analysis of the properly randomized trials. Meta-analyses based on individual patient data may also provide more information about treatment effects than the more usual overviews of grouped data, for they allow more detailed analyses—indeed, if they are really large then they may actually yield statistically reliable subgroup analyses of the effects of treatment in particular subgroups.

Conversely, even a perfectly conducted meta-analysis of an intervention with moderate effects on a major clinical outcome may not be reliable if the trials were all small. There are two reasons for this. First, when the true effect of an intervention is only moderate, most small trials will fail to reach statistical significance, and may be less likely to be published (or otherwise available) than the few with results that are misleadingly extreme. Hence, a meta-analysis consisting exclusively of small trials is particularly prone to bias. Secondly, the random errors may be too large to allow reliable interpretation. A meta-analysis that includes a total of only 100 deaths will have random errors about as great as a single trial with only 100 deaths. For these reasons small-scale evidence, whether from a meta-analysis or from one trial, is often unreliable and may well be found in retrospect to have yielded wrong answers. What is needed is large-scale randomized evidence; it does not matter much whether the totality of the evidence comes from a properly conducted meta-analysis of several trials or one properly conducted trial with such clear results that no further trials were done. The practical medical value of large-scale randomized evidence will be illustrated by a few examples.

**Examples of important results in the treatment of vascular disease that could have been reliably established only by large-scale randomized evidence**

**Definite result from a single very large trial: benefit from medium-dose aspirin for patients with acute myocardial infarction (with benefits among other types of patient indicated by meta-analyses of smaller trials)**

In the ISIS-2 trial, half of 17 000 patients with suspected acute myocardial infarction were allocated aspirin tablets (162 mg/day for 1 month, which virtually completely inhibits cyclooxygenase-dependent platelet activation) and half were allocated placebo tablets. Before 1988, when the ISIS-2 results were published, aspirin was not routinely used in the treatment of acute myocardial infarction, and no other major trial had (or has subsequently) compared aspirin with an untreated control group in cases of suspected acute myocardial infarction. However, the effects of 1 month of aspirin were so definite in ISIS-2 (804/8587 vascular deaths among those who were allocated aspirin vs 1016/8600 among those who...
so, so again a nonrandomized study might have had much less influence on medical practice than ISIS-2 did.

As a result, worldwide treatment patterns changed sharply when the ISIS-2 results emerged in 1988, and aspirin is now routinely used for the majority of emergency hospital admissions with suspected acute myocardial infarction not only in Europe and America but throughout Asia. In the United Kingdom, for example, two British Heart Foundation surveys found cardiologists reporting that routine aspirin use in acute coronary care had increased from under 10% in 1987 to over 90% in 1989. Worldwide, the annual number of patients with suspected myocardial infarction who would nowadays be given such treatment must be several million, and it is estimated that 1 to 2 million strokes, heart attacks, or vascular deaths. Small randomized trials and small meta-analyses of trials, or nonrandomized studies (however large), could not possibly have provided appropriately reliable evidence about such moderate risk reductions.

Fig. 2.3.3.3 Effect of administration of aspirin for 1 month on 35-day mortality in the 1988 ISIS-2 trial among 17000 patients with acute myocardial infarction. (Absolute survival advantage: 24\% lives saved per 1000 patients allocated aspirin, 2p=0.00001. The COMA trial in 46000 such patients has since shown aspirin plus clopidogrel to be slightly more effective than aspirin alone.)

In the ISIS-2 trial, aspirin significantly reduced the 1-month mortality, but it also significantly reduced the number of nonfatal strokes and nonfatal reinfarctions that were recorded in hospital. Combining all these three outcomes into ‘vascular events’ (i.e. stroke, death, or reinfarction), 10% of those who were allocated aspirin vs 14% of the controls suffered a vascular event in the month after randomization (Table 2.3.3.2)—an absolute difference of 40 events per 1000 treated (or, perhaps more relevantly, 40000 per million). The randomized trials of aspirin, or of other antiplatelet regimens, in other types of high-risk patients (e.g. a few years of aspirin for those who have survived a myocardial infarction or stroke) were not as large as ISIS-2, and so, taken separately, most yielded false-negative results. However, when the results from many such trials are combined, statistically definite reductions in ‘vascular events’ are seen (Table 2.3.3.2). Since such treatments do not appear to increase nonvascular mortality, all-cause mortality is also significantly reduced. More recently, a combination of aspirin and clopidogrel (which inhibits platelet activation through different pathways) has been shown to be slightly more effective than aspirin alone in acute myocardial infarction or acute coronary syndrome (Table 2.3.3.2).

The large-scale randomized evidence on anti-platelet drugs that is summarized in Table 2.3.3.2 has changed clinical practice worldwide, and may already have affected the treatment of hundreds of millions of patients in ways that, at low cost, have prevented millions of strokes, heart attacks, or vascular deaths. Small randomized trials and small meta-analyses of trials, or nonrandomized studies (however large), could not possibly have provided appropriately reliable evidence about such moderate risk reductions.

Definite result from a very large meta-analysis of trials: benefit from ‘adjuvant’ therapy with tamoxifen for patients with hormone-sensitive (ER-positive) ‘early’ breast cancer

By definition, in ‘early’ breast cancer all detectable deposits of disease are limited to the breast and the local or regional lymph nodes, and can be removed surgically. However, experience shows that undetectably small deposits of breast cancer cells may remain elsewhere that eventually cause clinical recurrence at a distant site, perhaps after a delay of several years, which is then usually followed by death from the disease. If the original tumour was ‘ER-positive’ (i.e. if the tumour cells were still expressing the oestrogen receptor protein) then the distant deposits of cancer cells that spread from it before it was removed may also be ER-positive, and may be continually stimulated by circulating hormones. Therefore, among women who have had breast cancer removed by surgery (or by surgery and radiotherapy), there have been many trials of ‘adjuvant’ daily treatment with tamoxifen, a drug that blocks the oestrogen receptor. Some involved only 1 to 2 years of treatment, some involved about 5 years, some compared 5 years vs 1 to 2 years and some that are still in progress compare 10 years vs 5 years of tamoxifen: in total, more than 100000 women have been randomized in several dozen such trials.

Taken separately, most of these tamoxifen trials have been too small to provide reliable evidence about long-term survival. However, if the results of all of them are combined in various ways, some very definite differences emerge: 1 to 2 years of tamoxifen is better than nothing, 5 years is better than 1 to 2 years, and 10 years may be better still for delaying or avoiding the recurrence of...
**Table 2.3.3.2** Summary results of (a) trials of aspirin (or other antiplatelet drugs), and (b) trials of adding clopidogrel to aspirin

<table>
<thead>
<tr>
<th>Type of patient</th>
<th>Study</th>
<th>Mean duration (total randomized)</th>
<th>Stroke, heart attack, or vascular death</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Antiplatelet vs control</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acute heart attack</td>
<td>ISIS-2</td>
<td>1 month (20 000)</td>
<td>10 (20 000)</td>
</tr>
<tr>
<td>Acute stroke</td>
<td>CAST and IST</td>
<td>1 month (40 000)</td>
<td>9 (40 000)</td>
</tr>
<tr>
<td>Previous heart attack</td>
<td>ATT</td>
<td>2 years (20 000)</td>
<td>13 (20 000)</td>
</tr>
<tr>
<td>Previous stroke/TIA</td>
<td>ATT</td>
<td>2.5 years (23 000)</td>
<td>18 (23 000)</td>
</tr>
<tr>
<td>Other high risk (e.g. angina, peripheral vascular disease)</td>
<td>ATT</td>
<td>1 year (20 000)</td>
<td>8 (20 000)</td>
</tr>
<tr>
<td>(b) Aspirin plus clopidogrel vs aspirin alone</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acute coronary syndrome</td>
<td>CURE</td>
<td>9 months (13 000)</td>
<td>9 (13 000)</td>
</tr>
<tr>
<td>Acute heart attack</td>
<td>CCS2-COMMIT</td>
<td>1 month (46 000)</td>
<td>9 (46 000)</td>
</tr>
</tbody>
</table>

Promising meta-analysis of small trials confirmed by large trials: benefit from fibrinolytic therapy in acute myocardial infarction

If a recent thrombus has just blocked a coronary artery, thereby causing acute myocardial ischaemia or infarction, fibrinolytic drugs (such as streptokinase or tissue plasminogen activator) can sometimes rapidly dissolve the thrombus, restoring the flow of blood and reperfusing the heart muscle. These drugs were first introduced into clinical research in the late 1950s, but the trials of fibrinolytic therapy for suspected acute myocardial infarction in the 1960s and 1970s were too small to be statistically reliable (none involved even 1000 patients). So, by the early 1980s the haemorrhagic side effects were obvious, the benefits had not been convincingly demonstrated, and such treatments were generally considered to be definitely dangerous, probably fairly ineffective, and hence inappropriate for routine coronary care. Although meta-analyses published in the mid 1980s of the previous small trials (which had involved a total of only 6000 patients in 24 trials) indicated a statistically definite benefit, they were not really believed by cardiologists and so such treatments were still not widely used.

The situation was saved by two large randomized trials, GISSI-1 and ISIS-2, which together involved about 30 000 patients (and by their aggregation with the seven other randomized trials that each involved more than 1000 patients, yielding a total of 60 000; see below). In ISIS-2, not only were patients randomly allocated to receive aspirin or placebo tablets as described earlier (Fig. 2.3.3.3), but also they were separately allocated to receive intravenous streptokinase or a placebo infusion. In this ‘factorial’ design (which allows the separate assessment of more than one treatment without any material loss in the statistical reliability of each comparison), one-quarter of the patients were allocated aspirin alone, one-quarter were allocated streptokinase alone, one-quarter were allocated both streptokinase and aspirin, and one-quarter were allocated neither (i.e. they were given placebo tablets and a placebo infusion). Streptokinase, like aspirin, produced a highly significant reduction in mortality, and the combination of streptokinase and aspirin was highly significantly better than either aspirin alone or streptokinase alone (Fig. 2.3.3.5). The results shown in Fig. 2.3.3.5 might suggest that there was no need to collect any more randomized evidence about fibrinolytic therapy, but this ignores the potential hazards of such treatment and the heterogeneity of patients. Taken separately, even ISIS-2 (the largest of these trials) was not large enough for statistically
reliable subgroup analyses, but when the nine largest trials were all taken together, they included a total of about 60,000 patients, half of whom had been randomly allocated fibrinolytic therapy. Those entering a coronary care unit with a diagnosis of suspected or definite acute myocardial infarction range from patients who are already in cardiogenic shock with low blood pressure and a fast pulse (half of whom will die rapidly) to those who have merely had a history of chest pain and no very definite changes on their ECG (of whom ‘only’ a small percentage will die before discharge). Fibrinolytic therapy often causes blood pressure to fall: should it be used in patients who are already dangerously hypotensive? It occasionally causes serious strokes: should it be used in patients who are elderly or hypertensive, and therefore already have an above-average risk of cerebral haemorrhage (or who have only slight changes on their ECG, and therefore have only a low risk of cardiac death)? Finally, if a coronary artery has been totally occluded for long enough, the heart muscle that it supplies will have been irreversibly destroyed: how many hours after the heart attack starts is fibrinolytic treatment still worth risking—3? 6? 12? 24?

These questions needed to be answered reliably before appropriate and generally accepted indications for and against such an immediately hazardous, but potentially effective, therapy could be devised. To address them, the main fibrinolytic therapy trialists collaborated in a systematic meta-analysis of the randomized evidence, based on individual patient data. On review of the 60,000 patients randomized between fibrinolytic therapy and control in trials of more than 1000 patients, some of the therapeutic questions were relatively easy to answer satisfactorily. For example, it appears that most of those whose ECG is still normal (or shows a pattern that indicates only a small immediate risk of death) can be left untreated, leaving open the option of starting fibrinolytic treatment urgently if their ECG changes suddenly for the worse over the
next few hours or days. Conversely, among those who already had ‘high-risk’ ECG changes when they were randomized, the absolute benefit of immediate fibrinolytic therapy was, if anything, slightly greater than is indicated by Fig. 2.3.3.5. Age, sex, blood pressure, heart rate, diabetes, and a previous history of myocardial infarction could not identify reliably any subgroup that would not, on average, have their chances of survival appreciably increased by treatment.

By contrast, the longer that fibrinolytic treatment for such patients was delayed the less benefit it seemed to produce. Among the 45 000 whose ECG showed definite ST-segment elevation or bundle-branch block, the benefit was greatest (about 30 lives saved per 1000) among those randomized between 0 and 6 h after the onset of pain (Fig. 2.3.3.6).

However, the mortality reduction was still substantial and significant (about 20 per 1000; 2p < 0.003) for the patients whose hospital admission had been delayed for some hours and who were therefore randomized 7 to 12 h after the onset of pain. Indeed, even if patients were randomized 13–18 h after the onset of pain, there still appeared to be some net reduction in mortality (about 10 per 1000, but not statistically definite). The regression line in Fig. 2.3.3.6 reinforces these separate subgroup analyses in a more reliable way. Yet, before these large trials it was forcefully, but mistakenly, argued that such treatments could not possibly be of any worthwhile benefit if given more than about 3 or 4 h after the onset of symptoms.

Such detailed inferences are difficult enough with large-scale properly randomized evidence, and would be impossible without it. Because of their unknowable biases (see above), nonrandomized database analyses are simply not a viable alternative to large-scale randomized evidence. Nor would randomization of ‘only’ several thousand patients have been sufficient. The availability of large-scale randomized evidence, in this case a meta-analysis involving about 6000 deaths among 60 000 patients, has been essential in determining which particular types of patient derive net benefit from fibrinolytic therapy.

Promising meta-analysis of small trials refuted by large trials: lack of significant benefit from magnesium infusion in acute myocardial infarction

In animal studies, infusion of a magnesium salt can limit the myocardial damage arising from sudden experimental blockage of a coronary artery. By the early 1990s, there was considerable optimism that a simple, inexpensive magnesium infusion might prove beneficial after acute myocardial infarction. Twelve small trials, involving between them a total of only about 2000 patients, had addressed this question, and their aggregated results indicated a highly statistically significant—but implausibly large—halving of risk (72/1199 deaths among those allocated magnesium vs 131/1191 among the controls, 2p < 0.00001). At this time some argued that such results constituted proof beyond reasonable doubt that magnesium was of sufficient value to justify its widespread usage without seeking further randomized evidence, but others remained sceptical, arguing that the apparent results were far too good to be true.

Two trials, one (LIMIT-2) involving 2000 patients and one (ISIS-4) involving 58 000, were then set up to test the possible effects of magnesium more reliably. The first yielded a moderately promising result (Fig. 2.3.3.7), indicating avoidance of about one-quarter of the early deaths, but with its 99% confidence interval including the possibility that magnesium had no beneficial effect on early mortality. The second (which had continued in spite of intense lobbying of its data monitoring committee to stop the trial), however, yielded a completely unpromising result, so that the overall evidence, by that time based on over 60 000 randomized patients, indicated no net effect on mortality.

Nevertheless, some cardiologists remained hopeful that magnesium might prove to be effective among specific subgroups. Accordingly the MAgnesiuM In Coronaries (MAGIC) trial subsequently randomized 60 000 patients, all of whom had received reperfusion therapy within the past few hours, to magnesium vs placebo, but this also found no evidence of any net benefit.
It is interesting to consider what this sequence of magnesium trial results (Fig. 2.3.3.6) might mean for those wishing to interpret other randomized evidence. Our interpretation is that if something seems too good to be true then it probably is—or, more formally, that big benefits are often much less plausible than moderate benefits. None of the 12 small trials had sufficient power to detect a moderate effect on mortality, and although their aggregated results indicated that mortality could be reduced by more than half, such an effect is too extreme to be plausible, and could be misleading even though it is highly significant. The LIMIT–2 trial then suggested that magnesium might reduce mortality by about a quarter, a result that is somewhat more plausible but not clearly significant. The success of the ISIS–4 and MAGIC trials in refuting the implausibly large benefit suggested by the 13 smaller trials reinforces our point that often, when trying to distinguish between the two medically realistic possibilities of a moderate effect or no effect, only large-scale evidence suffices. Even the LIMIT–2 trial, which recruited 2000 patients, was in retrospect too small. (Another important methodological point is that ‘random effects’ methods for meta-analysis can produce importantly wrong answers: applied to the 15 separate trials in Fig. 2.3.3.7, a standard ‘random effects’ meta-analysis yields a summary odds ratio of 0.67 (95% CI 0.52–0.85; 2p < 0.001), suggesting—clearly incorrectly—that magnesium reduces mortality by about one-third!)

**Trials in their epidemiological context: blood pressure, stroke, and heart disease**

Quantitative epidemiological evidence about the effects of long-term differences in risk factors such as blood pressure or blood cholesterol level can help in interpreting the results from trials of the effects of reducing these risk factors for only a few years. For example, appropriate meta-analyses of prospective observational epidemiological studies indicate that, throughout the range of usual systolic blood pressure in the populations studied (about 115–180 mmHg), a lower value is associated with a lower risk of ischaemic heart disease, with no apparent ‘threshold’ in this range below which the relationship reversed (Fig. 2.3.3.8). This analysis suggests that, in later middle age (60–69 years), 10 mmHg lower systolic blood pressure is associated with about 27% less mortality from ischaemic heart disease (and about 35% less stroke mortality; Prospective Studies Collaboration, data not shown).

By the mid 1990s, several trials had been conducted to determine whether a few years of blood pressure reduction in middle age reduces the risk of stroke and of coronary heart disease. Partly because of imperfect compliance, the mean difference in systolic blood pressure between the treatment and control groups in these trials was only about 10 mmHg. Even if such trial treatments would eventually produce about 27% less coronary heart disease after many years (as seen in observational studies), the effects seen within the 2 or 3 years that are available on average between randomization and death in a 5-year trial might well be somewhat smaller (perhaps only c.15%). But, considered separately, none of the trials recorded enough coronary heart disease events (or enough vascular deaths) for statistically reliable assessment of a 13% risk reduction.

For stroke, the trials provide direct and highly significant evidence that most, or all, of the risk reduction associated with 10 mmHg lower usual systolic blood pressure appears soon after the blood pressure is lowered (Fig. 2.3.3.9). In contrast, the significant reduction in coronary heart disease seen in the trials (16% SD 4, 95% CI 8–23%; 2p = 0.0001) seems to fall somewhat short of the difference of about 27% suggested by the observational evidence. However, the coronary heart disease reduction in the trials is still substantial and real (2p = 0.0001).

Taken together, Figs. 2.3.3.8 and 2.3.3.9 suggest that antihypertensive regimens that produce differences of much more than 10 mmHg systolic blood pressure will eventually reduce stroke by more than half and heart disease by more than a quarter. They also
suggest that the proportional risk reduction produced by a given absolute reduction in systolic blood pressure will be approximately independent of the initial systolic blood pressure.

**Results from large anonymous trials are relevant to real clinical practice**

A clinician is used to dealing with individual patients, and may feel that the results of large trials somehow deny their individuality. This is almost the opposite of the truth, for one of the main reasons why trials have to be large is just because patients are so different from one another. Two apparently similar patients may run entirely different clinical courses, one remaining stable and the other progressing rapidly to severe disability or early death. Consequently, it is a danger in too detailed an analysis of the apparent responses of small subgroups chosen for separate emphasis because of the apparently remarkable effects of treatment in these subgroups. Even if an agent brought no benefit, it would have to be acutely poisonous for it not to appear disproportionately beneficial in one or two such subgroups! Conversely, if an intervention really avoids an approximately similar proportion of the risk in each category of patient, it will, by chance alone, appear not to work in some category or categories of patient. The surprising extent to which this happens is evident from the example in Table 2.3.3.2. A large, anonymous trial will at least still help to answer the practical question of whether, on average, a policy of widespread treatment (except where clearly contraindicated) is preferable to a general policy of no immediate use of the treatment (except where clearly contraindicated). Moreover, without really large trials it is difficult to see how else many such questions relating to the effects of treatments on death or disability (or other major outcomes) are to be resolved reliably. Trials are at least a practical way of making some solid progress, and it would be unfortunate if desire for the perfect (that is, knowledge of exactly who will benefit from treatment) were to become the enemy of the possible (that is, knowledge of the average direction and approximate size of the effects of treatment in many large categories of patient).

**Further reading**

Antithrombotic Trialists’ Collaboration (2002). Collaborative meta-analysis of randomised trials of antithrombotic drug treatment (mean systolic blood pressure differences of c.10 mmHg for 5 years). Conventions are as for Fig. 2.3.3.7.

**Fig. 2.3.3.9** Reduction in the odds of stroke and coronary heart disease in all unconfounded randomized trials of antihypertensive drug treatment (mean systolic blood pressure differences of c.10 mmHg for 5 years). Conventions are as for Fig. 2.3.3.7.

Those few individuals who really stand to benefit from therapy. If any criteria (e.g., a short-term response to a non-placebo-controlled course of some disease-modifying agent) can be proposed that are likely to discriminate between people who will and will not benefit, then these can be recorded prospectively at entry and the eventual trial result subdivided with respect to them. However, there is a danger in too detailed an analysis of the apparent responses of small subgroups chosen for separate emphasis because of the apparently remarkable effects of treatment in these subgroups. Even if an agent brought no benefit, it would have to be acutely poisonous for it not to appear disproportionately beneficial in one or two such subgroups! Conversely, if an intervention really avoids an approximately similar proportion of the risk in each category of patient, it will, by chance alone, appear not to work in some category or categories of patient. The surprising extent to which this happens is evident from the example in Table 2.3.3.2. A large, anonymous trial will at least still help to answer the practical question of whether, on average, a policy of widespread treatment (except where clearly contraindicated) is preferable to a general policy of no immediate use of the treatment (except where clearly indicated). Moreover, without really large trials it is difficult to see how else many such questions relating to the effects of treatments on death or disability (or other major outcomes) are to be resolved reliably. Trials are at least a practical way of making some solid progress, and it would be unfortunate if desire for the perfect (that is, knowledge of exactly who will benefit from treatment) were to become the enemy of the possible (that is, knowledge of the average direction and approximate size of the effects of treatment in many large categories of patient).
overview of early mortality and major morbidity results from all randomised trials of more than 1000 patients. *Lancet*, 343, 311–22.


