The need for large-scale randomized evidence

C. Baigent
Clinical Trial Service Unit, Radcliffe Infirmary, Oxford OX2 6HE, UK

The reliable detection of moderate differences in major health outcomes that arise as a result of treatment requires large-scale randomized evidence (and the appropriate interpretation of this evidence once it has been generated). This may take the form of a single mega-trial or, exceptionally, a meta-analysis of many smaller randomized trials may provide worthwhile information. Small or non-randomized studies cannot generally be trusted to distinguish reliably between a moderate benefit, a moderate hazard, and a negligible difference in major outcomes. Simple design, streamlined data collection, and use of the ‘uncertainty principle’ to guide eligibility would all encourage the recruitment of larger samples in randomized trials. Future trials need to adopt these methods in order to detect any moderate improvements in major outcomes that may await discovery.

Keywords: randomized controlled trial, meta-analysis

The need to assess MODERATE differences in outcome

If a widely practicable treatment can achieve just a moderate reduction in a common cause of premature death or of serious disability, reliable recognition of this could prevent much suffering. Consider, for example, the case of aspirin, which is cheap, practicable, and widely available, and provides just a moderate degree of protection against the recurrence of myocardial infarction or occlusive stroke [1]. Worldwide, there are almost 10 million deaths a year from myocardial infarction or stroke [2], plus a comparable number of non-fatal (but often seriously disabling) episodes, and appropriately widespread use of aspirin for the secondary prevention of vascular disease could well avoid about 100,000 premature deaths annually.

Conversely, if a widely used treatment has no material effect on fatal or seriously disabling outcomes (or if an easy, inexpensive treatment is about as good as a more complex one) then reliable demonstration of this could avoid a lot of trouble, toxicity and expense in future medical practice. Thus clinical research into major outcomes needs to be able to distinguish reliably between a moderate benefit, a moderate hazard and a negligible difference in such outcomes. By contrast, large differences are so rarely encountered that clinical research strategies should generally not be centred on the search for them.

The need to be able to recognise moderate differences in the chiefly dichotomous outcomes that are of most relevance to patients (e.g. dead/alive, disabled/not) leads, in many cases, to the need for large-scale randomized evidence (and for appropriate analysis of it once it has been generated). The reasons are simple [3–5]: non-randomized methods cannot generally guarantee the avoidance of moderate biases, small-scale studies cannot guarantee the avoidance of moderate distortion by the play of chance, and even when large-scale randomized evidence is available (either in one or two mega-trials or in a large meta-analysis of many smaller trials) moderate distortions can easily be introduced by inappropriate analysis of it, especially if this involves unduly data-dependent emphasis on particular parts of the evidence—for example, on particular trials or on particular subgroups of patients.

How large a reduction in the risk of a major outcome would be plausible?

Enthusiasm for the theoretical and experimental foundations of a particular therapeutic approach often leads to exaggerated hopes for the effectiveness of that treatment in terms of its effect on major outcomes. These hopes may stem from dramatic effects on laboratory measures of efficacy, or on the types of surrogate outcomes that are commonly studied before drugs commence Phase III studies: an anti-cancer drug, for example, may shrink tumours impressively, or a novel antithrombotic may practically abolish experimental thrombosis. But only rarely do these large effects on surrogate endpoints translate into large effects on major clinical outcomes, and the overwhelming majority of recent improvements in outcome in the common cancers or in the major cardiovascular diseases have generally been only moderate in size. Relative risk reductions of only about 10 or 20% are therefore the most that should generally be anticipated (which might typically generate absolute differences in a major outcome of a few per cent at most), whereas relative risk reductions of 50% in the main outcome of interest* are implausible. Future studies aiming to assess

* For rare side-effects, however, there may be large proportional differences between one treatment and another, or between treatment and control. For example, some non-steroidal anti-inflammatory drugs increase the risk of gastrointestinal bleeding by 5-fold or more. Rare side-effects with such extreme relative risks can often be recognized reliably by non-randomized methods, and such relative risks are often best quantified in case-control or cohort studies.
the effects of treatments on major outcomes ought, therefore, to start with the premise that any risk reductions are unlikely to be large.

### Two fundamental requirements for the reliable assessment of moderate treatment effects

Any study whose main objective is to identify moderate treatment effects must ensure that any biases and random errors that are inherent in its design are substantially smaller than the effect which is to be measured. This limits the range of study designs that can be informative (Table 1) [6].

**Negligible biases**

If moderate differences are to be assessed then moderate biases must be avoided, and this often requires appropriate analysis of properly randomized evidence. Non-randomized study designs, in particular, cannot generally be guaranteed to exclude moderate biases, and are therefore practically useless for the identification of moderate treatment effects.

Even when studies are properly randomized (and, wherever possible, treatment allocation is concealed) moderate biases may be introduced during their analysis or interpretation. This can occur when patients are excluded after randomization, particularly when the prognosis of the excluded patients in one treatment group differs from that in the other (such as might occur, for example, if non-compliers were excluded after randomization). If there is, in reality, virtually no difference in outcome between two treatments, then the least biased assessment of treatment effect is that which compares all those allocated to one treatment vs all those allocated to the other (i.e. an ‘intention-to-treat’ analysis), irrespective of what treatment they actually received. Thus, whether analysing just one trial or a meta-analysis of many smaller trials, it is important to avoid post-randomization withdrawals.

A more important bias is that generated by unduly data-dependent emphasis on particular trials or on particular subgroups of patients. Although this is often defended as a strategy for understanding trial results in terms of disease mechanisms, it may lead to seriously misleading results. This is because reliable identification of categories of patient for whom treatment is particularly effective (or ineffective) requires surprisingly large quantities of data. Even if the real sizes of the effects of treatment do vary substantially among subgroups of patients, subgroup analyses are so statistically insensitive that they may well fail to demonstrate these differences. On the other hand, even when highly significant ‘interactions’ are found, they may be a poor guide to genuine differences among particular categories of patients—especially when such interactions have emerged after an overzealous examination of multiple subgroups. This might not matter but for the fact that such subgroup analyses are both widely reported and widely believed, which may lead to the inappropriate management of many thousands of patients worldwide. For example, in the large Italian trial, GISSI-1, comparing streptokinase vs control after acute myocardial infarction, streptokinase appeared to be beneficial only among patients without prior myocardial infarction (MI). Fortunately, however, the authors were circumspect about this finding [7]. This turned out to have been wise since a subsequent overview of all the large fibrinolytic trials showed that the benefits of fibrinolytic therapy after acute MI were similar irrespective of a history of myocardial infarction [8]. Many thousands of patients with a previous history of myocardial infarction might well have been denied fibrinolytic therapy, however, if this observed pattern in the GISSI-1 subgroups had been believed. In general it is preferable to emphasise the overall results of trials and to regard the results of post-hoc subgroup analyses with healthy scepticism.

This example also reinforces the importance, when evaluating the effects of a given treatment on major outcomes, of considering all the randomized evidence, preferably within a meta-analysis. Such meta-analyses help to avoid unduly data-dependent emphasis on especially striking results within particular trials, and hence provide a better guide to the true effects of treatments. Occasionally, however, when detailed information on individual patients is available within a meta-analysis that includes several thousand major outcomes (such as death [8] or recurrence of a major cancer [9]), it may then be feasible to identify particular groups of individuals in whom the benefits or hazards of treatment are especially large (whereas, in some meta-analyses the numbers are too small for even the main analyses to be statistically reliable, let alone the subgroups). Where it has been possible to establish cooperation between trialists before any of the trial results are known, publication of subgroup hypotheses beforehand [10] provides further protection against unduly data-dependent emphasis on particular results.

**Small random errors**

Whilst avoidance of moderate biases requires careful attention both to the randomisation process and to the analysis and interpretation of the available trial evidence, the avoidance of moderate random errors requires large numbers of events. Since major outcomes such as death may occur

---

**Table 1. Requirements for reliable assessment of MODERATE treatment effects** [6].

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negligible bias</td>
<td>(i.e. guaranteed avoidance of MODERATE biases)</td>
</tr>
<tr>
<td>— Proper RANDOMIZATION</td>
<td>(non-randomized methods cannot guarantee the avoidance of moderate biases)</td>
</tr>
<tr>
<td>— Analysis by ALLOCATED treatments</td>
<td>(i.e. an ‘intention-to-treat’ analysis)</td>
</tr>
<tr>
<td>— Chief emphasis on OVERALL results</td>
<td>(with no unduly data-derived subgroup analysis)</td>
</tr>
<tr>
<td>— Systematic OVERVIEW</td>
<td>(or ‘meta-analysis’) of all the relevant randomized trials</td>
</tr>
</tbody>
</table>

| Small random error | (i.e. guaranteed avoidance of MODERATE random errors) |
| — LARGE NUMBERS | (with minimal data collection since detailed statistical analyses of masses of data on prognostic features generally add little to the effective size of a trial) |
| — Systematic OVERVIEW | (or ‘meta-analysis’) of all the relevant randomized trials |
quite infrequently, really large numbers of patients often need to be studied before the results can be guaranteed to be statistically (and hence medically) convincing. For common diseases, the results of randomized trials may determine the management of many thousands (and perhaps even millions) of future patients. There is an obvious need, therefore, to ensure that they are reliable.

In general, the scale of randomized evidence that is necessary for the assessment of major outcomes is often not available even through a meta-analysis of all of the completed randomized trials, since too few major outcomes may have occurred—even in aggregate—for the results to be trustworthy. Instead, those who are planning to address the need for randomized evidence may have to design new randomized trials from scratch which, ideally, should aim to randomize sufficient numbers to generate reliable answers.

Randomized trials have become larger in recent years but still the very largest trials in most specialities do not exceed a few thousand patients. Even though such trials may seem ‘large’ to those who do the hard work, they may well not be large enough. Consider, for example, a widely practicable treatment among patients with acute MI that reduces in-hospital mortality from 10% to 8%. If just one million out of the five million such patients a year were to be treated, this would prevent about 20,000 deaths a year, which is a substantial achievement. But, such benefits would be surprisingly easy to miss. If, for example, just 2000 patients were to be randomized, 1000 to active treatment and 1000 to control, then we would expect about 80 deaths in the former and 100 in the latter, but even if exactly this result were to be seen it would not be conventionally significant (2P>0.1)—and, by chance alone, the results from such a trial might well be about 90 vs 90 deaths instead of about 80 vs 100, misleadingly suggesting the treatment to be completely useless. Even after randomizing 2000 patients, therefore, the real effects of this particular treatment might remain unclear [6]. What is needed, instead, is to randomize 20,000: this would allow the clear demonstration of a 20% reduction with 800 treated and 1000 control deaths (2P<0.0001), and somewhat smaller benefits which might still be worthwhile could also be detected.

Randomized trials can be large if they are kept simple

If trials aiming to assess the effects of treatments on major outcomes are to become substantially larger then as many as possible of the main barriers to rapid recruitment need to be removed. One of the most effective ways to guarantee the recruitment of small numbers of patients is to burden busy clinicians with obtaining large amounts of baseline information. The information recorded at entry can often be surprisingly brief, and should concentrate on those few clinical details which are of paramount importance (including major prognostic factors and those few variables that are thought likely to influence the effect of treatment). Likewise, complicated eligibility criteria, inappropriately complex consent procedures and extensive auditing of data all deter doctors from entering patients into studies. So, by limiting recruitment, they may result in trials that are too small to identify any moderate effects of treatments on major outcomes that may truly exist [6]. Furthermore, as well as discouraging recruitment, if trials are designed to be complex they are likely to incorporate a high cost per patient and, hence, they can only ever be small. Either way, complexity is rarely a virtue in trials designed to assess major outcomes, whereas simplicity can lead to the randomization of very large numbers of patients and to results which may lead to worldwide changes in practice within very short periods of time [1, 8, 11].

The ‘uncertainty principle’: ethicality, heterogeneity and maximal sample size

For ethical reasons, randomization can be contemplated only if both doctor and patient feel substantially uncertain as to which trial treatment is best. The ‘uncertainty principle’ maximizes the potential for recruitment within this ethical constraint. Simply stated, the ‘uncertainty principle’ is that patients can be entered only if the responsible physician is substantially uncertain as to which of the trial treatments would be most appropriate for this particular patient. No patient should be entered if the responsible physician and/or the patient are for any medical or non-medical reason(s) reasonably certain that one of the treatments that might be allocated would be inappropriate (in comparison with some other treatment that could be offered to the patient in the trial or outside it) [12].

This approach to randomization encourages heterogeneity in the resulting trial population and this, in large trials, may add substantially to the practical value of the results. Among the early trials of fibrinolytic therapy, for example, most of the studies had restrictive trial entry criteria which precluded the randomization of elderly patients, so that these trials contributed nothing towards answering the important clinical question of whether treatment was useful among older patients. Other trials that did not impose an upper age limit, however, did include some elderly patients, and were therefore able to show that age alone is not a contraindication to fibrinolytic therapy [8].

The ‘uncertainty principle’ ensures not only ethicality and clinically useful heterogeneity but is also a simple and practical scheme which encourages the randomization of large numbers of patients. There is scope, therefore, for many more trials to adopt this as the fundamental principle that determines who is eligible.

Can alternative study designs substitute for large-scale randomized evidence?

As the resources will never be available to design large, simple trials to address all the questions of clinical interest, it is reasonable to ask whether there are any circumstances when it might be possible to circumvent the need for large trials, either by using routinely collected observational data (sometimes referred to as ‘Outcomes research’) or perhaps by analysing previously published randomized trials (within meta-analyses).

Outcomes research

‘Outcomes research’ means various things to various people but, as commonly used, the term refers to the use of
Biases and random errors in small-scale meta-analyses

Since meta-analyses are appearing in medical journals with increasing frequency it is important that those responsible for the delivery of health care and the planning of future research are able to judge the reliability of such reviews. As for any epidemiological study, the three main sources of mistaken conclusions in meta-analyses are confounding, biases and random errors. In the context of a meta-analysis of properly randomized trials, confounding can arise when a ‘pure’ comparison of treatment A vs control, which aims to assess the effects of treatment A on a particular outcome, is ‘confounded’ by the routine co-administration with A of a second treatment (B, say) which might be expected to affect the incidence of the same outcome. This problem is easily avoided, however, by confining attention to unconfounded randomized trials where treatment A alone is compared against control (or when A plus other treatments is compared against those other treatments alone). The main problems that then remain are those of biases and random errors.

There are two main categories of biases which have the potential to affect the reliability of a meta-analysis: those that occur within individual trials; and those that relate to the selection of trials. Although more empirical research into the numerous biases that can occur within randomized trials would be valuable, it is clear from existing studies, for example, that inadequate treatment concealment can result in exaggerated estimates of treatment effect [15], and that the inappropriate post-randomization exclusion of particular patients is common [16]. The resulting biases may have unpredictable consequences for particular trials, however, so that generalizations about the impact of such biases are inappropriate.

A potentially more serious problem, however, results from selection biases associated with the process of identifying all relevant trials (or, on occasions, with the failure to include all such trials once they have been identified). Unfortunately, the subset of trials that are eventually published (and, hence, are conveniently available) is often a non-random sample of the unconfounded studies, since trials may well be more likely to be submitted for publication if their results are strikingly positive than if they are negative or null [17–20]. Such ‘publication bias’ can, along with other sources of bias, produce surprisingly convincing evidence of effectiveness for treatments that are actually useless [21]. The particular circumstances in which publication bias has contributed to producing misleading estimates of treatment are difficult to identify, and it is still more difficult to generalize about the likely size of such bias when it does occur. But the problem is likely to be particularly acute within small meta-analyses that contain less than a few hundred or so major outcomes, and consist chiefly of small trials. This is because the smallest trials are subject to the largest random errors, and are therefore capable of generating implausibly large effect estimates. If publication bias results in selected emphasis on some of the most promising of these trials, then the resulting summary odds ratios might well be particularly unreliable [22]. Hence, unless the particular circumstances of a small-scale meta-analysis suggest that publication bias is unlikely, it may be best to treat the results as no more than ‘hypothesis-generating’. On the other hand, a meta-analysis that in aggregate contains sufficient numbers of major outcomes to constitute ‘large-scale’ randomized evidence [1, 8, 9] is unlikely to be materially affected by publication bias and, provided there are no serious uncontrolled biases (see above) within individual component trials, is likely to be trustworthy.

Conclusions

The majority of medical interventions that have been subjected to systematic study in large-scale randomized trials have demonstrated at most only moderate effects on major outcomes such as death or serious disability. But just a moderate effect of treatment, if it can be demonstrated clearly enough for that treatment to be adopted widely, may well prevent substantial absolute numbers of premature deaths. Moreover, if a sustained search for moderately effective treatments is successful—as it has in myocardial infarction, for example—then the combination of two or three individually modest improvements in outcome may collectively result in substantial health gains. Unfortunately, there is no reliable alternative to large-scale randomized evidence for the identification of moderate effects on major outcomes, and for most of the important clinical questions where the existing trials do not provide such evidence the only medically and financially practical way to obtain it is to plan, design and conduct some large simple trials. If this is to happen then it will be important to encourage large-scale collaboration among disease specialists. This, for example, is already the case among those concerned with the treatment of children with acute lymphoblastic leukaemia in the UK, enabling over 90% of new cases to be entered into Medical Research Council trials [23]. Such collaborations are especially important, of course, among those concerned with the treatment of relatively uncommon diseases. Many disease specialists have already begun the important preliminary collaborative task of assembling the randomized evidence relevant to their
specialty within a Cochrane Review Group [24]. As well as providing useful information about treatments for which there is already reliable evidence of benefit, this may well be an efficient way of drawing attention to some of the many unanswered questions that are of relevance to patients. Moreover, some Cochrane Review Groups may then provide a focus for the large-scale collaborations that are going to be necessary if some large, simple trials are to be conducted in a wide range of seriously disabling diseases.

References


