Available online http://cvm.controlled-trials.com/content/2/1/002

Commentary

Fundamental deficiencies in the megatrial methodology

Bruce G Charlton

Department of Psychology, University of Newcastle upon Tyne, Newcastle upon Tyne; and Centre for Public Health Policy and Health Services Research, University of East London, London, UK

Correspondence: Bruce G Charlton MD, Department of Psychology, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, UK. Tel: +44 (0)191 222 6247; fax: +44 (0)191 222 5622; e-mail: bruce.g.charlton@newcastle.ac.uk

Received: 3 January 2001
Revisions requested: 12 January 2001
Revisions received: 17 January 2001
Accepted: 17 January 2001
Published: 30 January 2001

Curr Control Trials Cardiovasc Med 2001, **2**:2–7

(Print ISSN 1468-6708; Online 1468-6694)

Abstract

The fundamental methodological deficiency of megatrials is deliberate reduction of experimental control in order to maximize recruitment and compliance of subjects. Hence, typical megatrials recruit pathologically and prognostically heterogeneous subjects, and protocols typically fail to exclude significant confounders. Therefore, most megatrials do not test a scientific hypothesis, nor are they informative about individual patients. The proper function of a megatrial is precise measurement of effect size for a therapeutic intervention. Valid megatrials can be designed only when simplification can be achieved without significantly affecting experimental control. Megatrials should be conducted only at the end of a long process of therapeutic development, and must always be designed and interpreted in the context of relevant scientific and clinical information.

Keywords: epidemiology, history, megatrial, methodology, randomized trial

Introduction

Megatrials are very large randomized controlled trials (RCTs) – usually recruiting thousands of subjects and usually multicentred – and their methodological hallmark is that recruitment criteria are highly inclusive, protocols are maximally simplified, and end points are unambiguous (eg mortality). Megatrials have been put forward – especially by the 'evidence-based-medicine' movement' – as the criterion reference source of evidence, superior to any other method for measuring the effectiveness or effect size of medical interventions.

This aggrandizement of megatrials to a position of superiority is an error. I explore how it was that such a transparently ludicrous idea has gained such wide currency and explicate some of the fundamental deficiencies of the megatrial methodology which mean that – in most cases – megatrials are highly prone to mislead. Properly understood, the results of large, simplified, randomized trails can be understood *only* against a background of a great deal of other information, especially information derived from more scientifically rigorous research methods.

Reasons for the supposed superiority of megatrials

How did the illusion of the superiority of megatrials come about? There are probably three main reasons – historical, managerial, and methodological.

1. Historical

When large randomized controlled trials emerged from the middle 1960s, it was as a methodology intended to come at the end of a long process of drug development [1]. For instance, tricyclic and monoamine-oxidase-inhibitor antidepressants were synthesized in the 1950s, and their toxicity, dosage, clinical properties, and side effects were elucidated almost wholly by means of clinical observations, in animal studies, 'open', uncontrolled studies, and

MRC = Medical Research Council; NICE = National Institute for Clinical Excellence; NNT = number needed to treat; RCT = randomized controlled trial.

small, highly controlled trials [2]. Only after about a decade of worldwide clinical use was a *large* (by contemporary standards), placebo-controlled, comparison, randomized trial executed by the UK Medical Research Council (MRC), in 1965 – and even then, the dose of the monoamine-oxidase inhibitor chosen was too low. So, a great deal was already known about antidepressants *before* a large RCT was planned. It was already known that antidepressants worked – and the function of the trial was merely to estimate the magnitude of the effect size.

Nowadays, because of the widespread overvaluation of megatrials, the process of drug development has almost been turned upon its head. Instead of megatrials coming at the end of a long process of drug development, after a great deal of scientific information and clinical experience has accumulated, it is sometimes argued that drugs should not even be made available to patients until *after* megatrials have been completed. For instance, 1999 saw the National Institute for Clinical Excellence (NICE) delay the introduction of the anti-influenza agent Relenza® (zanamivir) with the excuse that there had been insufficient evidence from RCTs to justify clinical use, thus preventing the kind of detailed, practical, clinical evaluation that is actually a prerequisite to rigorous trial design.

It is not sufficiently appreciated that one cannot design an appropriate megatrial until one already knows a great deal about the drug. This prior knowledge is required to be able to select the right subjects, choose an optimal dose, and create a protocol that controls for distorting variables. If a megatrial is executed without such knowledge, then it will simplify where it ought to be controlling: eg patients will be recruited who are actually unsuitable for treatment, they will be given the trial drug in incorrect doses, patients taking interfering drugs will not be excluded, etc. Consequently, such premature megatrials will usually tend systematically to underestimate the effect size of a new drug.

2. Managerial – changes in research personnel

Before megatrials could become so widely and profoundly misunderstood, it was necessary that the statistical aspects of research should become wildly overvalued. Properly, statistics is a means to the end of scientific understanding [3] – and when studying medical interventions, the nature of scientific understanding could be termed 'clinical science' – an enterprise for which the qualifications would include knowledge of disease and experience of patients [1]. People with such qualifications would provide the basis for a leadership role in research into the effectiveness of drugs and other technologies.

Instead, recent decades have seen biostatisticians and epidemiologists rise to a position of primacy in the organization, funding, and refereeing of medical research – in other words, people whose knowledge of disease and patients in relation to any particular medical treatment is second-hand at best and nonexistent at worst.

The reason for this hegemony of the number-crunchers is not, of course, anything to do with their possessing scientific superiority, nor even a track record of achievement; but has a great deal to do with the needs of managerialism – a topic that lies beyond the scope of this essay [4].

3. Methodological – masking of clinical inapplicability by statistical precision

There are also methodological reasons behind the aggrandizement of megatrials. As therapy has advanced, clinicians have come to expect incremental, quantitative improvements in already effective interventions, rather than qualitative 'breakthroughs' and the development of wholly new treatment methods. This has led to demands for ever-increasing precision in the measurement of therapeutic effectiveness, as the concern has been expressed that the modest benefits of new treatment could be obscured by random error. Furthermore, when expected effect sizes are relatively small, it becomes increasingly difficult to disentangle primary therapeutic effects from confounding factors. Of course, where confounders (such as age, sex, severity of illness) are known, they can be controlled by selective recruitment. But selective recruitment tends to make trials small.

Megatrials appear to offer the ability to deal with these problems. Instead of controlling confounders by rigorous selection of subjects and tight protocols, confounding is dealt with by randomly allocating subjects between the comparison groups, and using sufficiently large numbers of subjects so that any confounders (including unknown ones) may be expected to balance each other out [5]. The large numbers of subjects also offer unprecedented discriminative power to obtain statistically precise measurements of the outcomes of treatment [6]. Even modest, stepwise increments of therapeutic progress could, in principle, be resolved by sufficiently large studies.

Resolving power, in a strictly statistical sense, is apparently limited only by the numbers of subjects in the trial – and very large numbers of patients can be recruited by using simple protocols in multiple research centres [6]. Analysis of megatrials requires comparison of the average outcome in each allocation group (ie by 'intention to treat') rather than by treatment received. This is necessitated by the absolute dependence upon randomization rather than rigorous protocols to deal with confounding [5]. So, in pursuit of precision, randomized trials have grown ever larger and simpler. More recently, there has been a fashion for pooling data from such trials to expand the number of subjects still further in a process called meta-analysis [7] – this can be considered an extension of the megatrial idea, with all its problems multiplied [8]. For instance, results of

meta-analyses differ among themselves, in relation to RCT information, and may diverge from scientific and clinical knowledge of pharmacology and physiology [9]

The problem is that 'simplification' of protocol translates into scientific terms as *deliberate reduction in the level of experimental control*. This is employed with good intentions – in order to increase recruitment, consistency, and compliance [5], and is vital to the creation of huge databases from randomized subjects. However, as I have argued elsewhere, the strategy of expanding size by diminishing control is a methodological mistake [10]. Reduced experimental control inevitably means less informational content in a trial. At the absurd extreme, the ultimate megatrial would recruit an unselected population of anybody at all, and randomize subjects to a protocol that would not, however, necessarily bear any relation to what actually happened to the subject from then on. So long as the outcomes were analysed according to the protocol to which the subject had originally been randomized, then this would be statistically acceptable. The apparent basis for the mistake of deliberately reducing experimental rigour in megatrials seems to be an imagined, but unreal, trade-off between rigour and size – perhaps resulting from the observation that small, rigorous trials and large, simple trials may have similar 'confidence interval' statistics [10]. Yet these methodologies are not equivalent: in science the protocol defines the experiment, and different protocols imply different studies examining different questions in different populations [5].

Assumptions behind the megatrial methodology

Megatrials could be defined as RCTs in which recruitment is the primary methodological imperative. The common assumption has been that with the advent of megatrials, clinicians now have an instrument that can provide estimates and comparisons of therapeutic effectiveness that are both clinically applicable and statistically precise. Widespread adoption of megatrials has been based upon the assumption that their results could be extrapolated beyond the immediate circumstances of the trial and used to determine, or at least substantially influence, clinical practice.

However, this question of generalizing from the average result of megatrials to individual patients has never been satisfactorily resolved. Many clinicians are aware of serious problems [11,12], and yet these problems have been largely ignored by the advocates of a trial-led approach to practice.

Extrapolation from megatrials to practice has been justified on the basis of several assertions. It has been assumed (if not argued) that high levels of experimental rigour are not important in RCTs because the randomization of large numbers of subjects compensates (in some undefined way) for lower levels of control. This is a mistaken argument based on a statistical confusion: large, poorly controlled trials may have a similar confidence interval to that in a small, well controlled trial (a large scatter divided by the square root of large numbers may be numerically equal to a smaller scatter divided by the square root of smaller numbers) – but this does not mean that the studies are equivalent [5]. The smaller, better-controlled study is superior. Different protocols mean a different experiment, and low control means less information. After all, if poor control were better than good control, scientists would never need to do experiments – control is of the essence of experiment.

Furthermore, it is routinely assumed that the average effect measured among the many thousands of patients in a megatrial group is also a measure of the probability of an intervention producing this same effect in an individual patient. In other words, it is assumed that the megatrial result and its confidence interval can serve as an estimate of the probability of a given outcome in an individual patient to whom the trial result might be applied.

This is not the case. Even when a megatrial population is representative of a clinical population (something very rarely achieved), when trial populations are heterogeneous average outcomes do not necessarily reflect probabilities in individuals. To take a fictional example: supposing a drug called 'Fluzap' shortens an illness by 5 days *if* that illness is influenza and *if* patients actually take the drug. Then suppose that the trial population also contains patients who do *not* have influenza (because of non-rigorous recruitment criteria) and also patients who (despite being randomized to 'Fluzap') do *not* take the drug – suppose that in such subjects, the drug 'Fluzap' has no effect. Then the *average* effect size for 'Fluzap' according to intention-to-treat analysis would be a value intermediate between zero and five – eg that 'Fluzap' shortened the episode of influenza by about a day. This trial result may be statistically acceptable, but it does not apply to any individual patient. The value of such a randomized trial as a guide to treatment is therefore somewhat questionable, and the mass dissemination of such a summary statistic through the professional and lay press would seem to be politically, rather than scientifically, motivated.

Confidence intervals – confidence trick?

The decline in scientific rigour associated with the megatrial methodology has been disguised by the standard statistical displays used to express the outcome of megatrials. Megatrials typically quote the statistic called the 'confidence interval' (CI) as their summary estimate of therapeutic outcome; or else quote the average outcome for each protocol and a measure of the 'statistical significance' of any measured difference between averages.

But although the confidence interval has been promoted as an improvement on significance tests [13], it has serious problems when used for clinical purposes, and is not a useful summary statistic for determining practical applications of a trial. The confidence interval describes the parameters within which the 'true' *mean* of a therapeutic trial can be considered to lie – with a quoted degree of probability and given certain rigorous (and seldom-met) statistical assumptions [14].

Clinicians need measures of outcome among *individual patients* in a trial, especially the nature and degree of variation in the outcome. The confidence interval simply does not tell the clinician what he or she needs to know in order to decide how useful the results of a megatrial would be for implementation in clinical practice. Average results and confidence intervals from megatrials conceal an enormous diversity among the results for individual subjects – for example, an average effect size for a drug is uninformative when there is huge variation between individuals.

When used to summarize large data sets, the confidence-interval statistic gives no readily apprehended indication of the scatter of patient outcomes, because it includes the square root of the number of patients as denominator (confidence interval equals standard deviation divided by square root of *n*) [15]. This creates the misleading impression that big studies are better, because simply increasing the number of patients will increase the divisor of the fraction, which will powerfully tend to reduce the size of the confidence interval when trials become 'mega' in size. Consequently, the confidence interval will usually reduce as studies enlarge, although the scatter of outcomes (eg the standard deviation) may remain the same, or more probably will increase as a result of simplified protocols and poorer control.

The exceptionally narrow 'confidence intervals' generated by megatrials (and even more so by meta-analyses) are often misunderstood to mean that doctors can be very 'confident' that the trial estimates of therapeutic effectiveness are valid and accurate. This is untrue both in narrowly statistical and broadly clinical senses. In fact, the confidence interval per se gives no indication whatsoever of the precision of an estimate with regard to the individual subjects in a trial. Furthermore, the narrowness of a confidence interval does not have any necessary relation to the reality of a proposed causal relation, nor does it give any indication of the applicability of a trial result to another population. Indeed, since the confidence interval gives no guide to the equivalence of the populations under comparison, differences between trial results may be due to bias rather than causation. [16].

So, narrow, nonoverlapping confidence intervals, which discriminate sharply between protocols in a statistical sense, may nevertheless be associated with qualitative variation between subjects such that a minority of patients are probably actively harmed by a treatment that benefits the majority [17].

Measures of scatter needed for clinical interpretation

It would be more useful to the clinician if randomized trials were to display their results in terms of the scatter of patient outcomes, rather than averages. This may be approximated by a scattergram display of trial results, with each individual patient outcome represented as a dot. Such a display allows an estimate of experimental control as well as statistical precision, since poorly controlled studies will have very wide scatters of results with substantial overlaps between alternative protocols. The fact that such displays are almost never seen for megatrials suggests that they would be highly revealing of the scientifically slipshod methods routinely employed by such studies.

If this graphic display of all results is too unwieldy even for modern computerized graphics, a reasonable numerical approximation that gives the average outcome with a measure of scatter is also useful – for example, the mean and standard deviation, or the median with interquartile range [14]. These types of presentation allow the clinician to see at a glance, or at least swiftly calculate, what range of outcomes followed a given intervention in the trial, and therefore (all else being equal, and when proper standards of rigour and representativeness apply) the probability of a given outcome in an individual patient.

While the confidence-interval statistic will usually give a misleadingly clear-cut impression of any difference between the averages of two interventions being compared, a mean and standard deviation reveal the degree of overlap in results. When the confidence interval relates to an interval scale, it may indeed be possible to use the confidence interval to generate an approximate standard-deviation statistic. This is done on the basis that the 95% CI is (roughly) two 'standard-error-of-the-mean' (SEM) values above and below the mean [15]. The SEM is the standard deviation divided by the square root of *n*. Therefore, if the difference between the mean and the confidence limit is halved to give the SEM, and if the SEM is multiplied by the square root of *n*, this will yield the approximate standard deviation. The above calculation may be a worthwhile exercise, because it is often surprising to discover the enormous scatter of outcomes that lie hidden within a tight-looking confidence interval. However, most megatrials use proportional measures of outcome (eg percentage mortality rate, or 5-year survival), and these measures cannot be converted to standard deviations by the above method, or by any other convenient means.

Confidence intervals therefore have no readily comprehensible relation to confidence concerning outcomes – which is the variable of interest to clinicians. What is required

instead of confidence intervals is a display, or numerical measure, of scatter that assists the practitioner in deciding the clinical importance that should be attached to 'statistically significant' differences between average results.

A false hierarchy of research methods leads to an uncritical attitude to RCTs

There is a widespread perception that RCTs are the 'gold standard' of clinical research (a hackneyed phrase). It is routinely stated that randomized trials are 'the best' evidence, followed by cohort studies, case–control studies, surveys, case series, and finally single case studies (quoted by Olkin [7]). This hierarchy of methods seems to have attained the status of unquestioned dogma. In other words, the belief is that RCTs are *intrinsically* superior to other forms of epidemiological or scientific study, and therefore offer results of greater validity than the alternatives.

To anyone with a scientific background, this idea of a hierarchy of *methods* is amazing nonsense, and belief in such a hierarchy constitutes conclusive evidence of scientific illiteracy. The validity of a piece of science is not determined by its method – as if gene sequencing were 'better than' electron microscopy! For example, contrary to the hierarchical dogma, individual case studies are not intrinsically inferior to group studies – they merely have different uses [18]. The great physiologist Claude Bernard pointed out many years ago that the averaging involved in group studies is a potentially misleading procedure that must be justified in each specific instance [19]. When case studies are performed as qualitative tests of a pre-existing explicit and detailed hypothetical model, they exemplify the highest standards of scientific rigour – each case serving as an independent test of the hypothesis [20,21]. Individual human case studies are frequently published in top scientific journals such as *Nature* and *Science*.

Validity is conferred not by the application of a method or technique, nor by the size of a study, nor even by the difficulty and expense of the study, but only by the degree of rigour (ie the level of experimental control) with which a given study is able to test a research question. Since megatrials deliberately reduce the level of experimental control in order to maximize recruitment, this means that megatrial results invariably require very careful interpretation.

NNT – not necessarily true

The assumption just mentioned is embodied in that cherished evidence-based medicine (EBM) tool, the comparison of two interventions in terms of the 'number needed to treat', or NNT [22]. The NNT expresses the difference between the outcomes of two rival trial protocols in terms of how many patients must be treated for how long in order to prevent one adverse event. For instance, comparing beta-blocker with placebo in hypertension may yield an NNT of 13 patients treated for 5 years to prevent one stroke.

However, the apparent simplicity and clarity of this information depends upon the clinical target population having the same risk–benefit profile as the randomized trial population. When trial and target populations differ and the trial population is unrepresentative of the target population, the NNT will be an inaccurate estimate of effect size for the actual patients whose treatment is being considered. For instance, an elderly population may be more vulnerable to the adverse effects of a drug and less responsive to its therapeutic effect, to the point where an intervention that produces an average benefit to the young may be harmful in the old.

On top of this, the patients in a megatrial population are *always* prognostically heterogeneous, because the methodology uses deliberately simplified protocols designed to optimize recruitment rather than control – and meta-analyses are even more heterogeneous [3,8]. In a megatrial that shows an overall benefit, it is very probable that while the outcome for some patients will be improved by treatment, other patients will be made worse, and others will be unaffected. What this means is that even a representative megatrial (and such trials are exceedingly uncommon) cannot provide a risk estimate of what will happen to individual patients who are allocated the same protocol. Trials on unrepresentative populations may, of course, be actively misleading. The NNT generated by a megatrial does not in itself, therefore, provide guidance for clinical management. The NNT is Not Necessarily True! [22].

Conclusion

Megatrials, like other kinds of epidemiological study, should be considered as primarily methods for *precise measurement* rather than a scientific method for generating or testing a hypothesis [10]. Precise measurements of the effect size of medical interventions such as drugs should be attempted *only* when a great deal is known about the drug and its clinical actions. When megatrials are conducted without sufficient background scientific and clinical knowledge, they will be measuring mainly artefacts. Unless – for instance – a trial is performed on pathologically and prognostically homogeneous populations, and uses well controlled management protocols, the apparent precision of the result is more spurious than real.

Megatrials have become an unassailable 'gold standard' in some quarters. And this situation has become self-perpetuating, since the results of megatrials have become de facto untestable. Since megatrials are not testing hypotheses, because they are merely measuring the magnitude of an effect, the result of a megatrial is itself not an hypothesis, and cannot be tested using other methods. A megatrial of, say, an antihypertensive drug measures the comparative effect of that drug under the circumstances of the trial. Assuming that no calculation mistakes have been made, this result of a megatrial is neither right nor wrong: it is just a measurement.

People often talk of megatrials as if they proved or disproved the hypothesis that a drug 'works'. Far from being the final word on determining the effectiveness of a therapy, this is a question that a megatrial is inherently incapable of answering. But once the error has been made of assuming that a statistical measurement can test a hypothesis, the mistake becomes uncorrectable, because the level of statistical precision in a megatrial is greater than that attainable by other methods.

In such an environment of compounded error, it should not really be a source of surprise that statistical considerations utterly overwhelm scientific knowledge and clinical understanding, and we end up with the lunacy of regarding statisticians and epidemiologists as the final arbiters of medical decision-making. Health care becomes merely a matter of managers providing systems to 'implement' whatever the number-crunching technocrats tell them is supported by 'the best evidence' [4]. The methodological deficiencies of megatrials make them ideally suited to providing an intellectual underpinning for that world of join-the-dots medicine which seems just around the corner.

References

1. Charlton BG: **Clinical research methods for the new millennium.** *J Eval Clin Pract* 1999, **5**:251–263.
2. Healy D: *The Antidepressant Era*. Cambridge, MA: Harvard University Press, 1998.
3. Charlton BG: **Statistical malpractice.** *J Roy Coll Physicians London* 1996, **30**:112–114.
4. Charlton BG: **The new management of scientific knowledge: a change in direction with profound implications.** In *NICE, CHI and the NHS Reforms: Enabling Excellence or Imposing Control?* Edited by Miles A, Hampton JR, Hurwitz B. London: Aesculapius Medical Press, 2000:13–32.
5. Charlton BG: **Mega-trials: methodological issues and clinical implications.** *J Roy Coll Physicians London* 2000, **29**:96–100.
6. Yusuf S, Collins R, Peto R: **Why do we need some large, simple randomized trials?** *Statistics Med* 1984, **3**:409–420.
7. Olkin I: **Meta-analysis: reconciling the results of independent studies.** *Statistics Med* 1995, **14**:457–472.
8. Charlton BG: **The uses and abuses of meta-analysis.** *Fam Pract* 1996, **13**:397–401.
9. Robertson JIS: **Which antihypertensive classes have been shown to be beneficial? What are their benefits? A critique of hypertension treatment trials.** *Cardiovasc Drugs Ther* **14**:357–366.
10. Charlton BG: **Megatrials are based on a methodological mistake.** *Brit J Gen Pract* 1996, **46**:429–431.
11. Julian D: **Trials and tribulations.** *Cardiovasc Res* 1994, **28**: 598–603.
12. Hampton JR: **Evidence-based medicine, practice variations and clinical freedom.** *J Eval Clin Pract* 1997, **3**:123–131.
13. Gardner MJ: S*tatistics with Confidence: Confidence Intervals and Statistical Guidelines*. London: British Medical Association, 1989.
14. Bradford Hill AB, Hill ID: *Bradford Hill's Principles of Medical Statistics*. London: Edward Arnold, 1991.
15. Kirkwood BR: *Essentials of Medical Statistics*. Oxford: Blackwell, 1988.
16. Charlton BG: **The scope and nature of epidemiology.** *J Clin Epidemiol* 1996, **49**:623–626.
17. Horvitz RI, Singer BH, Makuch, Viscoli CM: **Can treatment that is helpful on average be harmful to some patients? A study of the conflicting information-needs of clinical inquiry and drug regulation.** *J Clin Epidemiol* 1996, **49**:395–400.
18. Charlton BG, Walston F: **Individual case studies in clinical research.** *J Eval Clin Pract* 1998, **4**:147–155.
19. Bernard C: *An Introduction to the Study of Experimental Medicine.* New York: Dover, 1865; 1957 edition.
20. Marshall JC, Newcombe F: **Putative problems and pure progress in neuropsychological single-case studies.** *J Clin Neuropsychol* 1984, **6**:65–70.
21. Shallice T: *From Neuropsychology to Mental Structure.* Cambridge: Cambridge University Press, 1988.
22. Charlton BG: **The future of clinical research: from megatrials towards methodological rigour and representative sampling.** *J Eval Clin Prac* 1996, **2**:159–169.