Megatrials are based on a methodological mistake

BRUCE G CHARLTON

SUMMARY

Despite their prestige, megatrials are founded upon a methodological error. This is the assumption that randomization of very large numbers of subjects can compensate for deliberately reduced levels of experiemental control, but there is no trade-off between size and rigour. Randomized trials are not a 'gold standard' because no method is intrinsically valid — there are good and bad trials. Interpretation of megatrials is always difficult and requires considerable clinical and scientific knowledge. Three fundamental parameters should be considered when evaluating the applicability of a trial to clinical practice: rigour of design; representativeness of the trial population; and homogeneity of the recruited subjects.

Keywords: megatrials; randomized controlled trials; epidemiology.

IT may seem bizarre to suggest that megatrials — those much-vaunted, very large, usually multicentred, randomized controlled trials — are founded upon a mistake, yet the case seems conclusive. This mistake is one not of incompetence, but of applying a mathematical solution to a scientific problem. Therefore, the clinical interpretation of megatrials is a common instance of the phenomenon I have dubbed 'statistical malpractice'.[1]

Commentators have been misled into thinking that the nature of a randomized trial is primarily determined by the number of patients included,[2] but the primary quality that defines a randomized trial as 'mega' is not size but methodology. A *scientific* experiment (such as the randomized trials variously designated phase I, fastidious, explanatory or analytical)[3,4] will aim to achieve optimal control of interfering variables in order to test a hypothesis or measure a causal relationship. By contrast, a megatrial is characterized by the employment of a *deliberately lowered standard of experimental control* leading to the recruitment of clinically heterogeneous subjects to a study which estimates the association between protocol and outcome. Therefore, a megatrial should be interpreted as an epidemiological measurement technique, rather than a hypothesis-testing scientific method.[5]

Megatrials appear to offer unprecedented discriminative power to obtain precise measurements of the outcomes of treatment.[6] This statistical power is attained by recruiting very large numbers of patients, sometimes tens of thousands, randomly allocating them between therapeutic protocols, and comparing the average outcome from each group. This process yields an estimate of association between intervention and outcome with narrow 'confidence intervals' — an expression that is usually misunderstood to mean that doctors can be very 'confident' that the estimate is true. A further false interpretation is that precise statistics imply a similarly precise knowledge of prognosis in individual patients, and therefore, that megatrials provide highly specific and compelling guidance for clinical practice.

Both of these inferences concerning clinical applicability are incorrect, but nevertheless, they have led to megatrials being routinely overvalued. Megatrials are widely regarded as pre-eminent in a 'hierarchy' of research methodologies that range from anecdotal case reports and case series at the bottom, through case–control and cohort studies, to bigger and bigger randomized trials at the top. Randomized trials *per se* are conceived to be the epitome of medical 'science' — despite being non-scientific in aims and methods.[3-5] Such exaltation of *technique* to a benchmark of validity is absurd, rather like asserting that microscopy is the 'gold standard' of biology, without regard to the question being asked or the quality of the instrument.

The mistake

Although the methodology of megatrials seems flawed to many clinicians and scientists, given the sheer scale of current misinterpretation, it seems necessary to state the mistaken suppositions quite explicitly.

The crucial false assumption underlying megatrials is that a measurement can be made simultaneously more precise *and more valid* by reducing the rigour of its protocol in order to allow increased recruitment of patients. It is implicitly assumed that more numbers can compensate for less scientific rigour — presumably on the statistical basis that the measured 'confidence interval' is similar for both situations.

Therefore, megatrials employ a deliberately suboptimal experimental design in order to maximize recruitment and compliance, and in doing so, they are making the tacit supposition that a diminution in control can be *traded off* against an enhancement in numbers.[5] This assumption is wrong because the statistical precision of an estimate is uncorrelated with the truth of that estimate: estimates can be precise and false, or imprecise and true — and confidence intervals are irrelevant to the question.

A big crude trial is just a *different* experiment from a small rigorous trial, not a more precise variation of the same experiment. Megatrials and scientific trials have different aims, different designs and yield different information. There is no trade-off between size and control; the less rigorous experiment yields less information.[5]

How have megatrials achieved such an extraordinary status in the light of such manifest deficiencies? One reason is the confusion of statistical exactness with scientific knowledge,[1] but more important is the relentless pressure for ever more precise estimates of therapeutic effect, a demand arising principally from the requirements of quantitative modelling.[7,8] Typically, such a demand is reckless of the validity of estimates, and an estimate which is exact but probably wrong is thought to carry greater rhetorical weight than an estimate which is rough but right. It is perhaps unsurprising that it has been statisticians, epidemiologists and health economists (as well as accountants, managers, civil servants and politicians) who have been the most active propagandists for the use of megatrial results as a source of guidelines for medical practice. Clinical scientists remain sceptical.

The prestige of megatrials may also have arisen because of their superficial similarity to 'big science'. They are — after all — some of the largest, slowest and most expensive experiments

B G Charlton, MD, lecturer, Department of Psychology, University of Newcastle upon Tyne.

Submitted: 5 December 1995; accepted: 14 March 1996.

ever done in medicine, and are very hard work for hundreds of people. The achievement of overcoming formidable logistic difficulties with the aid of equally formidable amounts of money is an activity often mistaken for rigorous research.

The nature of clinical variation

However, the methodological root of megatrial mania is an excessive concentration on 'statistical' issues at the expense of the causal realities of biological and medical phenomena. In complex biological systems, such as the sick human, the major barrier to understanding is the extreme difficulty in locating the key factors and their causal inter-relationships. This is only possible when adequate levels of *control* are achieved.[9] Insufficient control of important interfering variables results in *bias* or *systematic error* (the distortion of measured comparisons owing to their unlikeness). Error is systematic when dissimilar variables or processes have been conflated and regarded as equivalent.[5]

By contrast, the fundamental methodological assumption of megatrials is that excessive random error is the most important source of variation in the clinical situation. Random differences ('noise') may conveniently be removed simply by averaging repeated measures, errors in one direction being balanced by errors in the other direction.

But truly random variation is very much the exception in medicine — being largely confined to the imprecision of physical measurement techniques. The main reason for unpredictable prognosis in response to treatment is that there are many *systematic* differences between patients and the causes operating upon them.[5] Therapeutic variations between patients (e.g. in a trial) are mostly the result of *qualitative* dissimilarities between those patients, such as differences in age, sex, social class, fitness, pathology, staging, severity, past medical history, coincident disease, concomitant medication, diet, local medical and nursing practices, and so on and on. Outcomes are different largely because inputs and processes are different. Such numerous and interacting instances of bias can be identified, understood and controlled only by a study design of the greatest possible rigour.

The megatrial methodology tries to circumvent this absolute requirement for rigour and is revealed as profoundly anti-scientific by its attitude towards experimental control.[10] Systematic errors (such as heterogeneity in age, severity and non-protocol treatment) are deliberately *allowed into the experiment* (to increase recruitment) then randomized between comparison groups. This indeed results in a 'fair comparison' of protocols, but at the cost of ensuring that these protocols are inadequate to control known and significant sources of bias, and as stated above, a poorly controlled experiment is less informative than a well-controlled one — after all, if poor control was a virtue, there would be no need to do experiments at all.

Similarities between megatrials and surveys

It is often stated that megatrials are 'pragmatic' rather than scientific investigations and that they are designed to provide guidance rather than explanations.[8] This view correctly emphasizes that megatrials are not in the business of conjecturing and testing causal hypotheses, but it implies that 'pragmatic' trials are straightforwardly applicable to clinical practice, whereas the problem of interpretation is actually extremely problematic.

Interpretation can be approached most validly by regarding a megatrial as essentially similar to a survey: the megatrial is a 'survey' of the outcome of randomly applying different protocols to a population.[10] Most megatrials, like surveys, fail to control all known sources of systematic error, and therefore, are concerned with correlations rather than unitary causes. Both methods also employ the same form of *inductive* inference: generalization of an estimate is valid (within statistical parameters) only when it extrapolates between like situations.[8] The use of inductive inference corresponds to the need for a study population to be a *representative sample* of the target population to which its results will be applied.

In an ideal megatrial, the study population should be representative of the target population: either a census or else a sample random with respect to relevant causal variables. The British trials of chemotherapy for acute lymphoblastic leukaemia (ALL) of childhood come close to this ideal of representativeness, being a near-census drawn from a relatively stable population of incident cases.[11] But most megatrials are performed on highly unrepresentative subjects: a biased sample drawn from an untypical population attending the trial centre.

And if a randomized trial is intended to be applicable to individual patients, then there is a further requirement that the trial subjects and clinical management be homogeneous (i.e. differing only randomly from one another with respect to all relevant variables that influence outcome). Again, a close approach to the ideal of homogeneity is attained by the paediatric UK ALL trials, in which diagnostic criteria are rigorous, and there is a high degree of uniformity of pathology, age, premorbid health and many other significant variables.[11] These paediatric trials have formed a model of iterative therapeutic progress. By contrast, two decades of methodologically simple megatrials in *adult* ALL –trials performed on a biased sample from a heterogeneous population, of widely varying pathology and prognoses, and subjected to a less tightly controlled protocol — have failed to achieve any significant advances in treatment efficacy.[12]

On formal methodological grounds, an estimate of therapeutic effect derived from a megatrial tells the clinician *nothing* about the experiences of the individual subjects in that trial: a moderate average improvement may summarize many combinations of benefits, harms and no effect among trial participants.[10] The mistake of applying to individual subjects a summary statistic from a heterogeneous population is well known (but nonetheless endemic) among epidemiologists as the 'ecological fallacy'.[8,13]

The hazards of interpreting megatrials: a job for clinical scientists

Problems with invalid megatrial estimates are now emerging in the medical literature. The systematic tendency for underestimation of differences between protocols ('null bias') owing to poor compliance and uncontrolled ancillary (non-trial) interventions has been described in relation to coronary heart disease.[14] This is merely one instance of the larger problem of lack of rigour. It has also been suggested that megatrials may have produced misleading or ambiguous estimates of therapeutic effect in a number of situations, including the impact of allogeneic marrow transplant on acute myeloid leukaemia, the role of warfarin anticoagulation in treating symptomless atrial fibrillation, and prophylaxis of stroke.[15-17] These problems are mainly caused by the heterogeneous and unrepresentative nature of megatrial study populations.

Scientists have no hesitation in ignoring sloppy experiments; epidemiologists should have a similarly ruthless attitude towards sloppy megatrials. However, the tendency (encouraged by the nonsense about randomized trials *as such* being a 'gold standard') is to assume that all megatrials yield valid inferences, constrained only by their statistical power. But from the preceding analysis, it is obvious that understanding the significance of the result of a megatrial is anything but straightforward. The bare summary statistic is worthless without a detailed knowledge of

the nature and magnitude of the sources of systematic error, and correction for these variables. And this requires knowledge of the identity and importance of specific interfering causes,[8,10] not merely a generalized interpretative approach that is based on statistical criteria and checklists.[18]

The mistake that underpins the megatrial methodology naturally affects all the second-order techniques that use megatrial results as their data. This undermines the validity of many instances of meta-analysis, cost-effectiveness or cost–utility studies, 'evidence-based medicine', decision analysis, and a host of other quantitative modelling strategies.[7,8]

It is becoming clear that the steeplechase after statistical precision at any price is clinically irrelevant at best and harmful at worst. The most clinically useful estimation is not 'statistical significance' (or its confidence interval clone), but a valid measure of the size of a therapeutic effect applicable to a clinically meaningful population.[19] Although attempts have been made to extract such information by secondary, subgroup analysis of megatrials,[17] this procedure is highly prone to introduce bias because the subgroups were not created by random allocation.[6] Clearly, there is no substitute for rigorous studies designed to answer specific questions.

The popular notion that megatrials are the gold standard — hence unimprovable — has hampered the development of alternative and superior approaches to clinical research and management. Instead of expending vast efforts and resources on performing ever larger and less controlled studies, a better strategy is to perform optimally controlled randomized trials on homogeneous clinical groups in the context of a fully recorded, census population cohort.[12] When applied to adult haematological malignancies, this 'population-based' approach has successfully overcome many of the problems of unrepresentative trials. Furthermore, the resulting estimates of therapeutic effect — although having less impressive confidence intervals than a megatrial — have the enormous virtue of providing information that is clinically relevant and potentially applicable to individual patients.[16]

Future historians of medicine will surely regard the current megatrial mania as a kind of South Sea Bubble of the 1990s. Big numbers, crude protocols and randomization are not a shortcut to useful knowledge.[1,8,20] If each megatrial is judged on its own merits in terms of the three fundamental parameters of homogeneity, rigour and representativeness, most studies will be rejected as misleading. At the very least, we should regard the interpretation of such deliberately suboptimal experiments as megatrials to be a fiendishly difficult task, requiring not just a reasonable degree of statistical sophistication and lots of data, but the highest possible levels of clinical and scientific expertise.

References

1. Charlton BG. Statistical malpractice. *J Roy Coll Phys London* 1996; **30:** 112-114.
2. Topol EJ, Califf RM. Answers to complex questions cannot be derived from 'simple' trials. *Br Heart J* 1992; **68:** 348-351.
3. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. *J Chron Dis* 1967; **20:** 637-648.
4. Feinstein AR. Randomized clinical trials. In: *Clinical epidemiology: the architecture of clinical research.* Philadelphia, PA: WB Saunders, 1985.
5. Charlton BG. Megatrials: methodological issues and clinical implications. *J Roy Coll Phys London* 1995; **29:** 96-100.
6. Yusuf S, Collins R, Peto R. Why do we need some large, simple controlled trials? *Stat Med* 1984; **3:** 409-420.
7. Feinstein AR. Clinical judgment revisited: the distraction of quantitative models. *Ann Int Med* 1994; **120:** 799-805.
8. Charlton BG. The scope and nature of epidemiology. *J Clin Epidemiol* 1996; in press.
9. Rosenberg A. *Instrumental biology, or the disunity of science.* Chicago, IL: Chicago University Press, 1994.
10. Charlton BG. Randomized clinical trials: the worst kind of epidemiology? *Nature Med* 1995; **1:** 1101-1102.
11. Chessells JM, Bailey C, Richards SM, MRC working party on childhood leukaemia. Intensification of treatment and survival in all children with lymphoblastic leukaemia: results of UK Medical Research Council Trial UKALL X. *Lancet* 1995; **345:** 143- 148.
12. Proctor SJ. Acute lymphoblastic leukaemia in adults: the case for a strategic shift in study approach. *Br J Haematol* 1994; **88:** 229-233.
13. Last JM. *A dictionary of epidemiology.* Oxford: Oxford University Press, 1988.
14. Woods KL. Megatrials and management of acute myocardial infarction. *Lancet* 1995; **346:** 611.
15. Sweeney KG, Pereira Gray D, Steele R, Evans P. Use of warfarin in non-rheumatic atrial fibrillation: a commentary from general practice. *Br J Gen Pract* 1995; **45:** 153-158.
16. Proctor SJ, Taylor PRA, Stark A, *et al.* Evaluation of the impact of allogenic transplant in first remission on an unselected population of patients with acute myeloid leukaemia aged 15–55 years. *Leukemia* 1995; **9:** 1246-1251.
17. Rothwell PM. Can overall results of clinical trials be applied to all patients? *Lancet* 1995; **345:** 1611-1614.
18. Fahy T, Hyde C, Milne R, Thorogood M. The type and quality of randomized controlled trials (RCTs) published in UK public health journals. *J Public Health Med* 1995; **17:** 469-474.
19. Rothman K. Lessons from John Graunt. *Lancet* 1996; **347:** 37-39.
20. Charlton BG. Practice guidelines and practical judgment: the role of megatrials, meta-analysis and consensus. *Br J Gen Pract* 1994; **44:** 290-291.

Acknowledgement
Thanks to Professor Steve Proctor and Dr Penny Taylor for haematological advice.