BioSocieties (2007), **2**, 115–127 © London School of Economics and Political Science
doi:10.1017/S1745855207005091

Commentary

A Defence of the Randomized Controlled Trial in Mental Health

Simon Wessely
Institute of Psychiatry, King's College London, De Crespigny Park, London SE5 8AF, UK
E-mail: S.Wessely@iop.kcl.ac.uk

> At the end of the day it is the faith of those NHS doctors most closely involved with the work of the Cancer Centre . . . which provides us with the most meaningful statistic of all. That when both patients and their doctors can testify to the intrinsic work of something they are unlikely to be proved wrong. (Anon., 1990)

My introductory quote for this article comes from a medical magazine, written by a doctor. The subject matter, *The Lancet* outcome study of patients treated at the Bristol Cancer Centre, is probably now only of historic interest. The quote, however, is pithy, eloquent, inspiring and flawed.

Doctors almost invariably think that what they are doing is right for the patient or they wouldn't be doing it. Mental health professionals, who are perhaps more reflective than some doctors, usually believe they are right as well. Most health-care professionals are well-meaning people who believe they do their best and that what they do is reasonably successful. They couldn't work otherwise. And usually they are right, because most illnesses, especially the ones we see in psychiatry, tend to improve. If not, they at least wax and wane, to give the illusion of improvement—like depression or asthma. From the patients' perspective, providing they meet a doctor or therapist who is nice, courteous and respectful, things are normally not too bad. Most patients report good outcomes because they may have got better anyway, and because they like their doctor or therapist. And, finally, all treatments seem to work better in acute single illnesses, as opposed to chronic, complex conditions occurring in patients with other multiple risk factors and disadvantages, so it is not surprising that we prefer to give our best treatments to those who seem most likely to benefit.

So how, then, can a doctor decide what really does work, and how can a patient know whether or not the treatment that the doctor recommends with such conviction stands a better than even chance of making him or her better? The answer is via the randomized controlled trial (RCT).

Simon Wessely is head of the department of Psychological Medicine at the Institute of Psychiatry, King's College London. His main research interests lie in the borders of medicine and psychiatry, amongst such areas as chronic fatigue syndrome, unexplained symptoms, and the so-called 'Gulf War syndrome'. He is a founding member of the Cochrane Depression, Anxiety and Neurosis Group, and also set up the Clinical Trials Unit at the Institute of Psychiatry, Maudsley and King's College Hospital. He is co-author with Brian Everitt of *The randomised controlled trial in psychiatry* (Oxford University Press, 2005). He is a strong believer in evidence-based medicine, at least compared to the opposite, but if he had his time again, he would like to have trained as a historian.

At the conference which in part gives rise to this special issue, I made a defence of the RCT, and I do so again in these pages. I do so not because I think that the RCT is the answer to all our problems, nor because I think it is without faults. I will discuss some of the limitations of the RCT later in the article, and I am very aware of the role of many other research methodologies in addressing the question of what works for whom. Indeed, my own CV contains rather more of these other methodologies than it does of RCTs, not least because of the difficulty and expense of conducting an RCT compared with other approaches. The RCT is not the pot of gold at the end of the evidence-based rainbow. But I remain convinced that the RCT comes closest to that elusive 'gold standard', and does indeed deserve the place it is given at the top of the hierarchy of evidence-based medicine. In this commentary I will attempt to explain why.[1]

Early developments of the RCT methodology

There is some dispute about when RCTs first entered medicine. According to Iain Chalmers, the first exposition of random allocation came from the Flemish physician Jean Baptiste van Helmont, writing in 1662, in which he advocated casting lots to decide which patients would receive blood-letting and which would not, and that the outcome measure would be the number of funerals in each group. However, there is no evidence that any contemporary physician accepted the challenge.

Anne Harrington (2006) recently drew attention to Benjamin Franklin's debunking of animal magnetism as the first placebo controlled trial, but, as she concludes, this belongs more to the history of medical scepticism than therapeutic assessment. Instead, many consider that the experiment performed by naval surgeon James Lind, in which he demonstrated the ability of citrus fruits to prevent scurvy, was the first practical controlled trial in medicine. In 1816, we find army surgeon Alexander Hamilton apparently using alternate allocation in a further attempt to ascertain the effectiveness or otherwise of blood-letting, although later historians have cast doubts on whether or not he ever carried out the experiments as reported.

One of the earliest accounts of the principles of randomization comes from the work of Thomas Balfour at the Royal Military Asylum in Chelsea in 1854. It is quoted by Chalmers (2001), and worth reproducing, since it conveys the essence of why randomization remains the best method of deciding if your treatment works. Balfour was unimpressed by the claims made of the ability of a homeopathic medicine to prevent scarlet fever in the orphan boys in his care:

> There were 151 boys of whom I had tolerably satisfactory evidence that they had not had scarlatina: I divided them into two sections, taking them alternately from the list, to prevent the imputation of selection. To the first section (76) I gave belladonna: to the second (75) I gave none: the result was that two in each section were attacked by the disease. The numbers are too small to justify deductions as to the prophylactic

1 An earlier version of this paper appeared as 'The randomised controlled trial', in M. Slade, & S. Priebe (Eds), *Choosing methods in mental health research* (London: Routledge, 2006: 85–98). Some of the material is also used in B. Everitt, & S. Wessely, *The randomised controlled trial in psychiatry* (Oxford University Press, 2005).

> power of belladonna, but the observation is good, because it shows how apt we are to be misled by imperfect observation. Had I given the remedy to all of the boys, I should probably have attributed to it the cessation of the epidemic.

As Chalmers comments, everything is there. The need for sound eligibility criteria (boys who had not yet had scarlet fever), the randomization (in this case by alternation), the problem of Type 2 errors (Balfour considered that his numbers were too small, and there remained a chance that belladonna did prevent scarlet fever, albeit very weakly), and the tangible risk of drawing an incorrect inference from uncontrolled data (the epidemic would appear to have been either over, or less virulent than previously thought, leading physicians to falsely believe that the relative absence of scarlet fever in the orphanage was due to belladonna).

Another milestone in the evolution of randomization can be partly claimed by psychiatry, since it concerns the work of William Fletcher, who demonstrated the role of polished rice in the aetiology of beri beri, and how this could be overcome by using uncured rice. He did so by randomly allocating (again by alternation) patients who were inmates of the 'lunatic asylum' in Kuala Lumpur.

By the 1930s there were numerous examples of clinical trials in which selection was determined by the toss of a coin, or alternative numbers. Why, then, is it traditional to describe the first fully randomized controlled trial as being the 1948 Medical Research Council (MRC) trial of streptomycin for tuberculosis? The answer is because of the role of the statistician, Austin Bradford Hill. The innovation he introduced in the 1948 trial was to tackle allocation concealment by introducing sealed envelopes. He did so specifically to prevent investigators having any possibility of influencing the selection of treatments. It is for this reason that the 1948 trial is justly celebrated as being the first of the modern genre of truly randomized controlled trials. The purist might point out that the MRC whooping cough trial actually preceded the tuberculosis trial, but it was the latter that reported first and has justly received the plaudits (Doll, 1998).

It is unclear who carried out the first truly randomized controlled trial specifically in psychiatry. David Healy (1997) gives four candidates:

- A placebo controlled randomly allocated trial of chlorpromazine for treating schizophrenia carried out in 1954 in Birmingham, UK, by the husband and wife team Joel and Charmain Elkes.
- Again in 1954, a trial performed by Linford Rees, who randomly allocated 100 anxious patients to either placebo or chlorpromazine.
- A trial undertaken at the Maudsley Hospital in London by David Davies and Michael Shepherd to study the use of reserpine for treating depression. This trial began in 1953 but reports of it did not appear until 1955. (Ironically, most modern psychiatrists who have heard of reserpine will associate it with producing, rather than alleviating depression!)
- Finally, during the same time period Mogens Schou and Eric Stromgren used a randomized trial to show the effectiveness of lithium as a treatment for mania.

Since trials take place over many years, perhaps it is invidious to try to label any one trial as the first in psychiatry. However, the first 'modern' large-scale trial, whose influence

continues to this day, is the 1965 UK Medical Research Council clinical trial for the treatment of depressive illness (MRC, 1965). The trial, which was conducted in three geographically dispersed regions within the UK, involved some 55 psychiatrists, recruited 269 patients with depression, randomized them to one of four treatment groups (two classes of antidepressant drug, ECT and a placebo), and then followed them for almost six months. It, more than any other, signalled a new era in the assessment of psychiatric treatments.

The particular need for RCTs in psychiatry

One of the principal changes in medical practice and culture during the last hundred years has been the increasing realization that it is not enough for a doctor to say that his or her treatment works, and nor is it enough for a patient to say likewise. These forms of anecdotal evidence, even if expanded into a series of anecdotes, dignified by the title of case series, are inadequate for the task.

There are, of course, exceptions to this. Clearly, if one takes a disease like bacterial meningitis, which was 100 percent fatal, and then introduces penicillin, after which it becomes almost 100 percent curable (assuming treatment is given in a timely fashion), a series of case reports is sufficient to establish benefit, and no one would even dream of doing a randomized controlled trial. The treatment of cardiac arrest is another example where case series evidence is enough. Sceptics of evidence-based medicine point out that there has never been a randomized controlled trial for the effectiveness of the parachute. However, this situation has never yet applied to psychiatry, and I suspect it never will. Why not?

First of all, many disorders in psychiatry improve spontaneously. Any treatment the patient may have received is therefore likely to be credited for this improvement by both patient and doctor. This accounts for much of the success of alternative therapies throughout history. We should not forget that generations of physicians would, in all honesty, have reported that bleeding was an effective treatment, and would be supported in this claim by those patients lucky enough to survive the intervention. Thus, in any disorder which is not universally fatal, anecdotal evidence will usually support any treatment claim.

Second, the process of spontaneous recovery is accentuated by what is called 'regression to the mean'. The symptoms of many disorders, such as depression, wax and wane. People tend to go to see the doctor when their symptoms are worse. The symptoms will improve over time, often due to the natural history of the condition, but the physician will sometimes falsely conclude that his or her intervention was responsible for this improvement, not taking into account that he or she is usually seeing the patient at their worst. For this reason regression to the mean is also called 'the physician's friend'.

The 'non-specific' effects of treatment, which overlap with, but are not the same as, the placebo effect are a third reason why case series are not a sufficient form of evidence in psychiatry (Hrobjartsson, 2001). The simple act of taking an interest in someone, listening to them, paying attention and giving them the expectation that you will do something about it is itself a powerful intervention. For that reason many charismatic doctors have, over the years, claimed great success for their particular treatment, whatever it may be, when the real intervention was provided by their character.

Fourth is selection bias. If one offers a treatment to 100 people, not all of them will accept. Often in psychiatry only a small proportion actually do. But this proportion is not random, and will almost invariably contain an over-representation of those with a good prognosis anyway. It may include those with more stable backgrounds, less severe illness, less comorbidity (such as drugs or alcohol), a more supportive home environment, a job to return to and so on. So, if someone gets better on Treatment A it may be that Treatment A actually works, or it may just be that those who accepted Treatment A would have a better outcome in any event. These factors, or confounders, that are associated both with the decision to accept treatment and with the outcome of the treatment, are alternative explanations for why the treatment seems to work.

The importance of randomization

So, if anecdote and the number of people successfully treated are a poor guide, how can we decide if a specific psychiatric treatment works or not? The answer involves randomization (Kleijnen *et al*., 1997; Wessely, 2001). Randomization deals with confounders by ensuring that they are distributed randomly (and hence without bias) between those who do, and those who do not, receive the treatment. It ensures that those who receive the treatment are not going to do better, or worse, because of some factor unrelated to the treatment. If patients have been randomly allocated to treatment or no treatment, then all of these factors should be equally distributed between the two groups, and any differences between the groups will either be due to the play of chance (and for that reason trials have to be reasonably large to eliminate that possibility) or because the treatment actually works.

Dealing with selection bias is the unique property of randomization. Controlled clinical trials also deal with other biases, like the use of placebos, blindness and rating scales to reduce observer bias, but randomization is the only way of overcoming selection bias. Its purpose is to ensure that like is being compared with like, and that hidden biases favouring one or the other arm of the trial have not crept in.

The beauty of randomization is that it not only deals with the confounders that you have thought of, but also those that you have not (Sibbald and Roland, 1998). Responses to a particular intervention, for example, are often better in females than in males. Gender, then, is a confounder: if one arm of a trial had more females than males, then the treatment tested would falsely appear to be superior. You eliminate the confounder by ensuring that the two arms have equal numbers of males and females. But what if you did not know about the gender confounder, and it only came out later? What if there are confounders you have never heard of, but the referees of your paper have? And what if there are confounders that are simply unknown at the present time? The only thing we can say with any confidence in psychiatry is that there is much we do not know about why some people respond better to any given treatment than others.

What happens if you do not randomize? The answer is simple. You are more likely to come up with the wrong answer. In a series of studies, it has been established beyond all doubt that when you do not randomize, all sorts of biases creep in (Antman *et al*., 1992; Chalmers *et al*., 1977, 1983; Kleijnen *et al*., 1997; Kunz and Oxman, 1998; Sacks *et al*., 1982, 1987; Schultz *et al*., 1994, 1995). And what these biases do is to systematically overstate

the effectiveness of the new treatment. Study after study comparing the results from evaluations of new treatments which do not include randomization find that these designs are far more likely to report that the new treatment works. Now it could be that, for some perverse reason, randomized controlled trials tend to be performed on weaker, less effective treatments, reserving the inferior research designs for the more powerful treatments. However, one can show the same even within randomized controlled trials: the better the design of the trial and the greater the protection from bias, the less the chance of showing that the new treatment works. Hence the importance given to what is called 'allocation concealment', or preventing the investigator from being able to influence the choice of treatment. We know that the greater the chance of the investigator being able to guess the next treatment, the more likely it is that the trial will be positive. Investigators have been known to do virtually anything to compromise randomization—holding 'opaque' envelopes to the light being merely one common trick—because they are convinced that they already know what is best for the patient (Schultz and Grimes, 2002). But what this actually shows is the unique power of the adequately concealed RCT to deliver unbiased information.

Randomization does not, of course, mean that you always get the 'right' answer. There are numerous examples of positive RCTs of treatments that later trials find ineffective. The example of St John's Wort for depression provides one instance (Shelton *et al*., 2001). The use of magnesium in the treatment of heart attacks provides a very famous non-psychiatric example. And one could argue that every positive trial of homeopathy provides another. There are numerous reasons why even properly randomized trials can still give incorrect answers, most often sample size and the play of chance. But randomization does protect against bias. You might be unlucky in a small trial and still get more treatment successes in the active group than the placebo group. Yet, provided that randomization was successful, this will not be due to bias, but chance, and the risk of this diminishes as the sample size increases. Bias, on the other hand, is not affected by sample size. A large biased study is even more dangerous than a small biased study, because people are more likely to be taken in by the number of noughts in the '*p*' value. All it shows is that these results did not occur by chance alone—it does not protect you from bias. In general one can say that large treatment effects in small trials are inherently less believable and more likely to be due to some violation of the principles of the RCT than small treatment effects in large trials: 'Moderate (but worthwhile) effects on major outcomes are generally more plausible than large effects' (Collins *et al*., 1996).

It is not for nothing that RCTs come at the top of the hierarchy of knowledge—a position first accorded them nearly 30 years ago because of their unique ability to deal with bias (Byar, 1978). And because bias, in all its shapes and sizes, is the single biggest enemy of all attempts to determine if our treatment (as opposed to our charm, luck or the natural history of illness) really does work, then RCTs are indeed the King or Queen of assessment techniques.

It is worthwhile, however, to briefly take note of the kinds of questions the RCT methodology can and cannot answer. The RCT addresses the question: does treatment A do more good than harm (or vice versa) than treatment B in condition C? It does *not* tell you why a treatment might work, although the use of placebo conditions can often shed much light on processes and is widely used in psychological experiments. But that is an additional benefit from some RCTs, not a prime reason for their existence. The RCT does *not* tell you

that treatment A will work on patient B; it can only tell you that, on balance, treatment A is more likely to do good than harm in a series of patient Bs. The RCT does *not* tell you that treatment A works on condition D, if that was not the focus of the original trial. Nor does it tell you if treatment A works on patient E, if patient E systematically differs from the patients in the original trial. But what the RCT *does* tell you, and uniquely so, is whether or not the benefits of treatment outweigh the risks, and if so, by how much. All treatments have side effects—there is no such thing as an effective intervention without side effects. What the RCT does is assess the balance between risk and benefit.

Arguments against RCTs in psychiatry

There have been a number of well-reasoned criticisms of the use of randomized controlled trials, particularly, but not only, in the field of mental health. For example, taking one voice from many, Silberschatz articulates the principal arguments against RCTs in psychiatry from the perspective of a psychotherapist (Persons and Silberschatz, 1998). For him the important questions are: what is bothering the patient? What do they hope to achieve? Why have they not achieved that? The argument continues that manualization, deemed essential in psychological treatment trials to enable another clinician to be able to repeat the intervention later and to ensure that the therapy is replicable, removes the heart of psychological treatment—empathy, therapeutic alliance and so on. What is lost, it is claimed, is the essential individual nature of psychological treatments. People are different, problems are different and therefore, the argument goes, treatments should be different. Dehue (2002) goes beyond the world of psychotherapy to argue that because history, culture and networks matter (an indisputable observation), RCTs are not a rational research strategy because they deny such interconnections (a *non sequitur*).

How can we counter such arguments against trying to evaluate psychiatric treatments scientifically via RCTs? It is of course true that people are different, but this applies across medicine. A hundred years of writing on the 'art of medicine', the recent growth of 'narrative-based medicine', and the seemingly endless critiques of the limitations (or at least the perceived limited scope) and the limited success of narrowly oriented biomedicine, show that, across the entire medical profession, no one should seriously dispute the importance of understanding the individual (see White, 2005).

But if that was all there was, if every patient was indeed unique and every problem without precedent, then medicine in general and psychiatry in particular would come to a full stop. If there were no commonalities between patients, and no identifiable general patterns in particular groups of patients, then there would be no purpose in medical education, or any purpose in clinical experience and training. It is these shared factors that permit clinicians to draw on what they have learnt from both their training and their experience, and then use this acquired knowledge to assess and understand the specific patient now requiring their attention. After all, an intelligent clinician does not treat every person as a completely unique entity; rather, we classify patterns and information to be able to apply hard-won knowledge about similar people encountered in the past to the person at hand.

And it is the existence of patterns of disease that makes clinical trials viable. Having observed some phenomenon previously in a patient population of interest—be it a certain

cancer, a particular behaviour, a biochemical abnormality or an emotional reaction—means there is something that might form the basis for a clinical trial. The systematically acquired information that results can be used to help future patients, without forgetting that what is truly unique about a patient (and so cannot be studied in a clinical trial) still has to be taken into account when caring for the patient, and for this the treating clinician will often need large amounts of intuition, experience and empathy.

Another argument against RCTs in mental health is that psychiatric disorders are too complex. True, psychiatric disorders are frequently not straightforward, and psychiatric patients often display challenging and complex behaviours that might at first sight appear incompatible with the tightly controlled demands of most clinical trials. Broad categories such as depression and schizophrenia hide several sub-groups whose boundaries are imperfectly delineated. Many psychiatric patients have more than one diagnosis. What use is it studying those rare patients in whom depression does not coexist with other disorders, such as anxiety or substance abuse, when in 'real life' these so often go together? And is it really possible to recruit members of 'difficult' patient populations and to maintain them in a trial according to the stringent requirements of the trial protocol?

Complications of diagnosis and patient complexity can both be difficult challenges faced by psychiatric trialists, but neither provides *fundamental* objections to the use of RCTs in psychiatry or elsewhere. Comorbidity, for example, may affect generalization, if the index trial was performed on an unusually 'pure' sub-group of patients, but the validity of the data is unaffected. And trials can be (and have been) conducted, and conducted to a high standard, in populations and situations that might seem insuperable to the faint-hearted. Schizophrenia and substance abuse, for example, does not seem an auspicious subject for an RCT, since patients with both problems ('dual diagnosis' in the jargon) are sometimes seen as 'unascertainable, unconsentable, untreatable and untrackable'. But a research group in Manchester, UK, performed just such a trial to good effect (Barrowclough *et al*., 2001). Again, it might seem impossible to carry out randomized trials in violent forensic patients, yet there is a seminal trial in which 321 mentally disordered offenders were randomly assigned to either be released or put on outpatient compulsory treatment (Swartz *et al*., 2001).

It has also been argued that interventions in mental health are simply too complex to be reduced to the simplicities of a clinical trial. It is certainly the case that, in mental health, we seem to have a vested interest in making things more complex than necessary. Diagnostic issues in psychiatry, for example, can become something of a fetish, and, taken to extremes, can undermine the inherent simplicity of the clinical trial. Few clinicians really care, for example, about the sub-divisions of somatoform disorders, or whether someone has dysthymia or double depression. And psychiatrists use far too many rating scales to measure far too many things in their trials, increasing the chances of false positive findings; as an Oxford, UK, group of trialists note 'many trials would be of much greater scientific value if they collected 10 times less data on 10 times more patients' (Collins *et al*., 1996). An analysis of trials on the Cochrane Schizophrenia Database found that over 640 different rating scales had been employed (Gilbody *et al*., 2002; Thornley and Adams, 1998). The use of a large number of outcome measures is driven by the fear of missing something that might be 'clinically significant' even if that 'something' was not the primary reason for carrying out the study. But any advantages of such an approach are usually outweighed by the

disadvantages, in particular those of multiple testing and loss of simplicity both in analysis and in understanding of results.

Yet another criticism concerns the generalizability of RCTs. Critics point out that many clinical trials take place in 'pure' populations, free from all forms of comorbidity, where participants are keen to attend follow-ups, happy to take medication, etc., with the consequence that the results are not considered relevant to the vast majority of the population who *do* suffer from comorbidity and who are, in general, reluctant to do any of the things mentioned. Likewise, prognostic features of patients in clinical trials may vary, even within trials, and it is certainly true that one cannot assume that, because a treatment has been successful in a well-conducted clinical trial, the results will apply to all patients with the same diagnosis (Rothwell, 1995).

This is indeed a powerful argument for more pragmatic trials, and it has to be admitted that the issue of generalizability is perhaps the most cogent criticism of the RCT as currently undertaken (McKee *et al.*, 1999). But note the rider, 'as currently undertaken'. The fault lies not with the principles of the randomized clinical trial, but simply the way such trials are often conducted at present. The answer is not for psychiatry to turn its back on the RCT, but to push for larger, simpler trials, and to lobby against the increasing bureaucratization of the clinical trial that stands in the way of achieving these objectives.

The RCT, of course, is also subject to manipulation. The act of selecting 'good' patients for trials is sometimes forced on investigators by regulatory bureaucracy, or may reflect the simple truth that such patients are easier to recruit. Occasionally it may reflect attempts to over-emphasize the efficacy of a new treatment because, as already outlined, well-behaved patients usually do better whatever treatment they are given. Finally, recent examples of selective publication of RCTs, usually within the pharmaceutical sector, are indefensible—not because the trials themselves are suspect (they usually are not), but because selective publication undermines the process of balancing harm versus benefit by introducing bias. However, these examples of malpractice do not represent a fundamental challenge to the status of the RCT, any more than detecting doctors who are less than competent or less than honest undermines the legitimacy of the medical enterprise.

Using RCTs in future mental health research

There are numerous instances in which the evidence produced by RCTs has proved itself superior to other forms of assessment. If we had not carried out clinical trials we would still be giving insulin coma to schizophrenics. Back when I was a 'proper' doctor and not a psychiatrist, the standard treatment for septicaemic shock, was high-dose steroids. We now know, because of randomized controlled trials, that more people die when you give them steroids than when you do not. Likewise, at the time I qualified, the treatment of cerebral malaria was high-dose steroids, but trials showed that this killed more people than it cured. When I was a cardiology Senior House Officer, I worked on coronary care, where we used to give a drug called lignocaine, a local anaesthetic agent, to people whose ECGs showed plenty of ventricular ectopics. It 'worked' because it did indeed suppress ventricular ectopics, but, again, trials showed that more people died from being given lignocaine than not. How could that have been shown except by a clinical trial with random allocation of

treatment? If one of my patients, to whom I had just given lignocaine, died, then I could always reassure myself with the thought that 'This is a coronary care unit, people here have had heart attacks, and lots do die.' It is only in a randomized trial that you can actually see that your treatment may be doing more harm than good.

To illustrate my point let me cite a classic example from the mental health literature: the debriefing controversy. Most people will be familiar with the concept of single-session psychological debriefing. This is a fairly structured procedure, in which a mental health professional carries out an intervention with people, either individually or in groups, very shortly after they have been exposed to some form of adversity. The procedure involves some element of telling the story of the event, asking how people felt emotionally both now and during the event, and teaching about likely further emotional reactions over time. Its purpose, enthusiastically proclaimed by its protagonists, is to prevent later psychiatric disorder such as Post-traumatic Stress Disorder, PTSD.

In our contemporary culture, the arrival of what the media inevitably call 'trained counsellors' has become as much a part of the theatre of disaster as that of the emergency services. It has become part of the social recognition of disaster, and our collective desire that 'something must be done' (Gist, 2002). And it seems to be very sensible. What harm could possibly come from talking to someone who has been exposed to a trauma? 'Better out than in' is now very much the fashion. Who could possibly think this not a good idea? Very few, judging by the number of indications for stress debriefing—in a quick literature search I recently found over 56 different scenarios in which stress debriefing is used or advocated.

Yet, does it work? Even to ask the question is to invite ridicule from some quarters. Early attempts by several investigators in the field to mount an RCT were blocked because some ethical committees felt that it was unethical to deny debriefing to disaster victims. The aficionados of debriefing, and there are many, meanwhile claim that there is no need for such trials since the evidence already exists: 'the experiences of over 700 CISM teams in more than 40,000 debriefings cannot be ignored, especially so when the majority of reports are extremely positive … [N]umerous studies have already shown positive results … prov[ing] the clinical effectiveness beyond reasonable doubt' (Mitchell and Everly, 2003).

Unfortunately, the opposite is true. The randomized controlled trials of debriefing are overwhelmingly negative. In a Cochrane meta-analysis, the Peto odds-ratio for short-term psychological distress is firmly anchored around unity. What is more, the two studies with the highest-quality scores and the longest follow-up show a significant increase in the risk of PTSD in those receiving debriefing (Wessely *et al*., 2000), a finding confirmed by a different *The Lancet* meta-analysis (van Emmerik *et al*., 2002).

Armed with this information, we can now come up with many possible reasons for the ineffectiveness and possible harm of debriefing. Psychologists might argue that it exposes people to the risk of retraumatization, without providing any subsequent therapy. Certainly trials that involve several sessions—and of cognitive behavioural therapy rather than debriefing—do provide more encouragement. Sociologists consider the professionalization of distress, and the extent to which debriefing impedes the normal ways in which we deal with adversity—talking to our friends, family, GP, the vicar and so on. However, the point is that only randomization could have given this information, and overcome the problems of

regression to the mean, high satisfaction (as opposed to efficacy—very different things) and multiple confounding. Without these trials it would have proven impossible to question the wisdom of debriefing.

So how should we use RCTs in future mental health research? The short answer is more often. The longer answer is more often, with more patients, and with fewer measures. We must make psychiatric trials 'bigger' (larger numbers of patients), 'simpler' (fewer outcome measures, for example) and more 'life-like' (in psychiatry, perhaps more so than any other discipline, the case for more pragmatic trials that reflect real clinical practice is compelling). But the problem of size, or power, remains the biggest challenge. When one has conditions of major public health importance, which include most psychiatric disorders, even modest treatment effects may have a major impact on populations. Yet, almost invariably, mental health trials are so small that they can only detect major treatment effects, which are often inherently implausible.

Take depression as an example. We know that both the tricyclics and the SSRIs are effective in management. But which is better? And what do we mean by 'better' anyway? We can agree that should one class of drugs be, say, 50 percent better (however defined) than the other, then this group would immediately become the treatment of choice and the results would represent a dramatic breakthrough in treatment. Even a 25 percent improvement in outcome from one class of antidepressants over the other would be of considerable importance, and indeed still be close to being a 'dramatic breakthrough'. But since depression is a very common problem worldwide (the World Bank analysis predicts that it will be the second most common cause of disability across the world by 2020 and, while that might be slightly tendentious, there is no denying its public health importance) even a 10 percent improvement produced by one class of drugs over the other would be a very worthwhile benefit.

There have been over a hundred trials that compare tricyclics and SSRIs, so presumably we should by now know the answer to this question. But we do not, and the reason is simple: the trials were too small. Hotopf and colleagues analysed 121 trials that compared tricyclics 'head to head' with SSRIs (Hotopf and Normand, 1997). Many of the trials were sufficiently large to be able to detect that SSRIs were about 50 percent better in improving outcome than tricyclics; none of course did, and such a quantum leap in efficacy was always improbable. Less than a dozen could have detected a 20 percent difference. And if the differences were 10 percent (perhaps the most realistic possibility) then not a single trial could have come anywhere near detecting what would still be an important improvement in the management of depressed patients.

I began with a quote and I end with one. It is from Richard Horton, the editor of *The Lancet*, who, although offering a critical look at modern trials, nevertheless conveys the continuing central importance of the randomized trial in promoting better health care:

> All health-care professionals directly or peripherally involved in clinical trials need to recommit themselves to explaining, proselytising, promoting, understanding, encouraging, studying, protecting, strengthening, and reflecting on the clinical trial process. (Horton, 2001)

References

Anon. (1990). Flawed study's mixed effects. In *Hospital Doctor*, 18 October.

Antman E., Lau J., Kupelnick B., Mosteller F., & Chalmers T. (1992). A comparison of results of meta-analyses of randomized controlled trials and the recommendations of clinical experts. *Journal of the American Medical Association, 268*, 240–248.

Barrowclough C., Haddock G., Tarrier N., Lewis S.W., Moring J., O'Brien R. et al. (2001). Randomized controlled trial of motivational interviewing, cognitive behavior therapy, and family intervention for patients with comorbid schizophrenia and substance use disorder. *American Journal of Psychiatry, 158*, 1706–1713.

Byar D. (1978). Sound advice for conducting clinical trials. *New England Journal of Medicine, 297*, 553–554.

Chalmers I. (2001). Comparing like with like: Some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments. *International Journal of Epidemiology, 30*, 1156–1164.

Chalmers T., Matta R., Smith H., & Kunzler A. (1977). Evidence favouring the use of antidepressants in the hospital phase of acute myocardial infarction. *New England Journal of Medicine, 297*, 1091–1096.

Chalmers T., Celano P., Sacks H., & Smith H. (1983). Bias in treatment assignment in controlled clinical trials. *New England Journal of Medicine, 309*, 1358–1361.

Collins R., Peto R., Gray R., & Parish S. (1996). Large-scale randomized evidence: trials and overviews. In D. Weatherall, J. Ledingham, & D. Warrell (Eds), *Oxford Textbook of Medicine*, 21–32. Oxford: Oxford University Press.

Dehue T. (2002). A Dutch treat: Randomized controlled experimentation and the case of heroin maintenance in the Netherlands. *History of the Human Sciences, 15*, 75–98.

Doll R. (1998). Controlled trials: The 1948 watershed. *British Medical Journal, 317*, 1217–1220.

van Emmerik A., Kamphuis J., Hulsbosch A., & Emmelkamp P. (2002). Single session debriefing after psychological trauma: A meta-analysis. *The Lancet, 360*, 741–742.

Gilbody S., Wahlbeck K., & Adams C. (2002). Randomized controlled trials in schizophrenia: A critical perspective on the literature. *Acta Psychiatrica Scandinavica, 105*, 243–251.

Gist R. (2002). What have they done to my song? Social science, social movements and the debriefing debates. *Cognitive and Behavioral Practice, 9*, 273–279.

Harrington A. (2006). The many meanings of the placebo effect:Where they came from, why they matter. *BioSciences, 1*, 181–193.

Healy D. (1997). *The antidepressant era*. Cambridge, MA: Harvard University Press.

Horton R. (2001). The clinical trial: Deceitful, disputable, unbelievable, unhelpful, and shameful—what next? *Controlled Clinical Trials, 22*, 593–604.

Hotopf M., Lewis G., & Normand C. (1997). Putting trials on trial—the costs and consequences of small trials in depression: A systematic review of methodology. *Journal of Epidemiology and Community Health, 51*, 354–358.

Hrobjartsson A., & Gotzsche P. (2001). Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. *New England Journal of Medicine, 344*, 1594–1602.

Kaptchuk T. (1998). Intentional ignorance: A history of blind assessment and placebo controls in medicine. *Bulletin of the History of Medicine, 72*, 389–433.

Kleijnen J., Gotzsche P., Kunz R., Oxman A., & Chalmers I. (1997). So what's so special about randomisation. In A. Maynard,, & I. Chalmers (Eds), *Non-random reflections on health services research: On the 25th anniversary of Archie Cochrane's Effectiveness and Efficiency*, 93–106. London: BMJ Publishing Group.

Kunz R., & Oxman A. (1998). The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *British Medical Journal, 317*, 1185–1190.

McKee M., Britton A., Black N., McPherson K., Sanderson C., & Bain C. (1999). Interpreting the evidence: Choosing between randomised and non-randomised studies. *British Medical Journal, 319*, 312–315.

MRC (Medical Research Council) (1965). Clinical trial of the treatment of depressive illness. *British Medical Journal, i*, 881–886.

Persons J., & Silberschatz G. (1998). Are results of randomized controlled trials useful to psychotherapists? *Journal of Consulting and Clinical Psychology, 66*, 126–135.

Rothwell P. (1995). Can overall results of clinical trials be applied to all patients? *The Lancet, 345*, 1616–1619.

Sacks H., Chalmers T., & Smith H. (1982). Randomized versus historical controls for clinical trials. *American Journal of Medicine, 72*, 233–240.

Sacks H., Berrier J., Reitman D., Ancona-Berk V., & Chalmers T. (1987). Meta-analyses of randomized controlled trials. *New England Journal of Medicine, 316*, 450–455.

Schultz K., & Grimes D. (2002). Allocation concealment in randomised trials: Defending against deciphering. *The Lancet, 359*, 614–618.

Schultz K., Chalmers I., Grimes D., & Altman D. (1994). Assessing the quality of randomization from reports of controlled trials published in obstetrics and gynaecology journals. *Journal of the American Medical Association, 272*, 125–128.

Schultz K., Chalmers I., Hayes R., & Altman D. (1995). Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatments effects in controlled trials. *Journal of the American Medical Association, 273*, 408–412.

Shelton R.C., Keller M., Gelenberg A., Dunner D.L., Hirschfeld R., Thase M.E. *et al.* (2001). Effectiveness of St John's wort in major depression—A randomized controlled trial. *Journal of the American Medical Association, 285*, 1978–1986.

Sibbald B., & Roland M. (1998). Why are randomised controlled trials important? *British Medical Journal, 316*, 201.

Swartz M.S., Swanson J., Hiday V.A., Wagner H.R., Burns B.J., & Borum R. (2001). A randomized controlled trial of outpatient commitment in North Carolina. *Psychiatric Services, 52*, 325–329.

Thornley B., & Adams C. (1998). Content and quality of 2000 controlled trials in schizophrenia over 50 years. *British Medical Journal, 317*, 1181–1184.

Wessely S. (2001). Randomised controlled trials: The gold standard? In C. Mace, S. Moorey, & B. Roberts (Eds), *Evidence in the balance*, 46–60. Hove: Routledge.

Wessely S., Bisson J., & Rose S. (2000). A systematic review of brief psychological interventions ('debriefing') for the treatment of immediate trauma-related symptoms and the prevention of post-traumatic stress disorder. In M. Oakley-Browne, R. Churchill, D. Gill, M. Trivedi, & S. Wessely (Eds), *Depression, anxiety and neurosis: Module of the Cochrane database of systematic reviews*. Oxford: Update Software.

White P. (Ed.) (2005). *Biopsychosocial medicine: An integrated approach to understanding illness*. Oxford: Oxford University Press.