

EDITORIAL

## Methodological aspects of randomized controlled trials of psychotherapy in primary care<sup>1</sup>

Randomized controlled trials (RCTs) of psychological interventions are still much less frequently published than drug trials. To have four papers (Bower *et al.* 2003; Proudfoot *et al.* 2003; Simpson *et al.* 2003; Gabbay *et al.* 2003) within a single issue of a leading psychiatry journal is indeed unusual. Together these four major papers provide evidence of the benefits and limitations of psychological interventions for common mental disorders.

Bower *et al.* (2003) report a systematic review and meta-analysis of counselling in primary care. This treatment approach has been used increasingly in the UK in recent years, but evidence of its value has been scanty. Bower *et al.* find evidence of improvement in short-term outcome compared to usual general practice care, but with a modest effect not apparent in the longer term. In keeping with this Simpson *et al.* (2003) report little evidence of benefit in a controlled trial of psychodynamic counselling in chronic depression.

Proudfoot *et al.* (2003) report a controlled trial of a different approach, a computerized interactive multimedia cognitive-behavioural program, in patients with depression or anxiety. Improvement compared with treatment as usual was found on a range of measures, at 2 months and up to 6 months. Shorter, less expensive approaches to psychological treatments have been widely sought, and are necessary if treatments are to be widely applicable in primary care.

Gabbay *et al.* (2003) report an analysis from a large controlled trial of cognitive-behavioural analysis and non-directive counselling. They examine an aspect of communication, agreement between patients, general practitioners and therapists regarding baseline problem formulations, and the factors contributing to it. Mutual agreement predicted outcome although weakly and they conclude that agreement that the core problem is psychological may be more important than the detail.

These four valuable papers together make important contributions and the authors are to be congratulated on them. The studies can also be used to illustrate the challenges researchers must overcome in this difficult field. This editorial will focus on methodological aspects.

Mainert (1986) provides guidelines on the key features of an RCT, namely, that the study will include a well-defined, representative sample of sufficient size to yield power to detect clinically significant differences between groups; that treatment and control conditions will be specified sufficiently to allow replication; that the sampling frame and the outcome measures will be justified and that the latter will be identified *a priori*; that measurement bias will be avoided or controlled, and that analysis of outcome data will be by intention to treat (i.e. if a subject is randomized their data will be analysed). Although these principles were initially devised for drug treatment trials; they are often used as a standard against which to measure the quality of psychological treatment studies. However, there are anxieties about the unquestioning application of the drug metaphor to psychological treatment studies. In what follows, we highlight some key points in the design and interpretation of RCTs of psychological interventions, drawing on the published papers as examples.

### STUDY DESIGN – EFFICACY OR EFFECTIVENESS?

The starting point of a robust RCT is a careful and accurate formulation of its aims. This applies equally well to psychological interventions and drug trials, even though the former are often more

<sup>1</sup> Address for correspondence: Professor Jan Scott, PO Box 96, Department of Psychological Medicine, Institute of Psychiatry, De Crespigny Park, London SE5 8AF.

complex, both conceptually and practically, than the latter. Bower and colleagues (2003) undertook a systematic review as to whether counselling was more effective in the short-term than usual treatment (treatment as usual – TAU) in primary care. As King (1995) highlighted, it is easier to define what counsellors do rather than what counselling is. Given the non-directive nature of the intervention, it is not surprising Bower *et al.* (2003) found it particularly hard to identify papers that specified treatment effect(s) on specific measures, such as symptom change, sufficiently to discriminate them.

Study designs can be divided into two broad categories – efficacy studies and effectiveness studies (Streiner, 2002*a*). Efficacy studies aim to test whether an intervention can give benefit under ideal circumstances. These RCTs aim to recruit homogeneous samples, often with a restricted range of diagnoses made according to research criteria. Such studies are often designed to maximize differences between the interventions being tested. The research question is essentially ‘can this intervention work?’. Sample attrition is relatively unimportant except as a general source of bias in the results. Because of these design features, efficacy studies are not necessarily helpful in deciding whether an intervention would work in day-to-day practice.

By contrast, effectiveness studies evaluate ‘does this intervention work in a real clinical setting?’. Effectiveness studies are sometimes called pragmatic trials. To answer the question adequately, effectiveness studies must test interventions in circumstances that resemble as closely as possible real clinical settings. This involves setting broad inclusion criteria so that patients recruited will resemble those routinely encountered in clinical practice. Ideally, it involves using as therapists clinicians who work in the practice setting, rather than providing specially trained research therapists. In clinical practice as in research, patients drop out of treatment. So the research question extends to ‘does this intervention have benefits, even allowing for the people who do not complete the intervention?’. This can only be answered if all the patients entered into the study are correctly accounted for at the end of the study, and their data comprehensively included. Hence an absolute requirement of effectiveness studies is intention-to-treat analysis, including appropriate management of (inevitable) missing data. Increasingly, journal editors are insisting the papers describing RCTs include information meeting agreed standards of data reporting (defined in the Consolidated Standards of Reporting Trials – CONSORT – statement) (Moher *et al.* 2001). We support the use of CONSORT in psychotherapy studies and would wish to see this standard applied to future publications.

Managing missing data can be particularly problematical in psychological interventions. A commonly used method is described as last-observation-carried-forward (LOCF). Here, missing data points are filled with the last data point available for that individual. Although widely used, this method, like all others, is at risk of producing misleading results in some circumstances (Streiner, 2002*b*), particularly because psychological interventions usually occur against a background of symptoms changing naturally over time. Finding more appropriate methods to replace missing data in psychotherapy trials is a major challenge. It is therefore important to note that Proudfoot *et al.* (2003) give a clear statement of both their rationale and strategy for data management.

Confusing efficacy with effectiveness, or confounding these within the design of a study, are common and inevitably give rise to misunderstanding. Indeed some criticisms of RCTs in psychotherapy are undermined because they have focused exclusively on efficacy studies (e.g. Seligman, 1996; Persons & Silberschatz, 1998), and have failed to acknowledge the existence of effectiveness trials. Among the studies in this issue, that of Simpson and colleagues (2003) has the features of an effectiveness trial. However, there is one difference between the methods used in this study and routine clinical practice. To reflect clinical practice, the general practitioner should have referred all patients to the study. As described in the paper, the recruitment method was changed to increase the sample size (see next section). The authors noted that the patients recruited by the two different methods were comparable. It is, however, always possible that the two patient groups might have been distinguished by important factors not related to the basic variables compared. For example, it can be that doctors decide against referring patients whose past histories suggest either that

they are unlikely to benefit from the intervention, or are likely to improve over time without intervention.

## STUDY SAMPLES

A common feature of all the studies in this issue is that the clinical populations are heterogeneous. Most did not meet or were not assessed on diagnostic criteria for any specific mental disorder and participants were identified as 'cases' or selected on the basis of scores above an arbitrarily determined cut-off point on screening instruments that measure symptom severity. Even in the study by Proudfoot *et al.* (2003), where subjects underwent a second screening assessment; nine different diagnostic subgroups were identified. This approach has many advocates as the sample can clearly be seen to reflect the populations attending general practice. It does also make the management of an RCT and the interpretation of the findings more complicated. For example, in some studies, slow rates of referral via general practitioners or other aspects of study design, meant that the researchers had to re-think their recruitment strategy and many potential participants were then approached directly from the waiting room at the general practice. The time and effort required by researchers in psychotherapy RCTs in primary care is not always acknowledged and it is particularly true that negative results often fail to be published. We therefore commend the way that Simpson *et al.*'s study has sought to publish their findings and also the thoughtful and complete description of their results. This aids comparisons on a number of parameters between this negative outcome study and other RCTs (Paykel *et al.* 1999; Keller *et al.* 2000) that demonstrate the benefits of other psychological interventions for chronic depression.

Finally, a common question when results are presented at conferences is who refused to participate and why? Few journals can afford to offer the researchers the space to detail all the reasons why individuals refused to participate. It was noteworthy that the study by Proudfoot and colleagues (2003) of the use of a computerized intervention package lost 100 potential subjects (30% of those eligible) because these individuals refused to complete a computerized version of self-rated screening instrument. It would be fascinating to know if this time efficient approach to assessment actually biased the sample towards those already at ease with the technology employed in the study. These examples are not chosen to undermine the importance of the findings of the papers published but simply to highlight why even high quality RCTs of any treatment for psychological problems in primary care may raise issues of interpretation.

## OUTCOMES

The complexity of psychological interventions, and the complexity of the problems which they aim to treat, warrant the use in RCTs of multiple outcome measures. However, especially in effectiveness trials, it is widely regarded as good practice to identify *a priori* a single primary outcome measure. In a placebo-controlled drug trial, it is usually possible to identify a single measure appropriate to the trial's aims. In psychological interventions, this becomes considerably more difficult; especially when two different psychological treatments are being compared or the researchers wish to capture subjective and observer perspectives. This is even more problematic in counselling or psychodynamic psychotherapy, where the focus is usually not usually the types of outcome measured in drug trials, such as depressive symptom severity. Thus, it is entirely appropriate to measure the efficacy of cognitive therapy or drugs for depression using the Beck Depression Inventory (BDI) as the primary outcome measure. But is reduction in BDI scores equally applicable as a primary outcome measure in counselling, when the term encompasses several different psychological approaches (Bower *et al.* 2003; Simpson *et al.* 2003)? Smith *et al.* (1980) previously suggested that several factors associated with rating scales (collectively referred to as reactivity) affect estimates of psychotherapy outcome. The key elements of reactivity are: the similarity between therapy goals and the measure itself; the degree to which the measure can be influenced by the client or therapist; and the degree of 'rater awareness' in the assessment process. There is a modest but notable correlation between reactivity and the effect size (about 0.2). For the same reasons, different outcome measures cannot

necessarily be taken as equivalent. From a logical and statistical standpoint we can only really compare like with like (Roth & Fonagy, 1996), but it may be necessary to combine studies with varying outcomes together in a meta-analysis. It could be argued that the resultant Forrest plots would allow the assessment of significant heterogeneity. Bower *et al.* (2003) found evidence for modest benefits from counselling in primary care (effect sizes of  $<0.3$  are generally regarded as weak). However, it could be that the effects and the observed homogeneity of Forrest plots reflect the use in the studies of outcome measures not strongly related to the actual effects of the intervention. An important message of this meta-analysis was how few RCTs were available on the benefits of counselling and how heterogeneous the approaches of the researchers were to the basic problem of what constitutes benefit.

Another problem to which effectiveness studies are particularly prone is the difficulty of demonstrating the overall effectiveness of a treatment across mixed diagnostic groups where the conditions involved may have very different natural histories. Outcomes cannot be divorced from the natural history of a specific condition (Roth & Fonagy, 1996), but RCTs in primary care that evaluate treatments in highly selected samples will not find support from clinicians struggling with mixed or ill-defined psychological problems. For better informed decisions about implementation of interventions in routine clinical practice, estimates of financial as well as (or instead of) specific clinical outcomes are increasingly incorporated into psychological therapy RCTs (for example Guthrie *et al.* 1999). However, such health economic outcomes are really only meaningful in studies designed to reflect routine clinical practice as accurately as possible i.e. effectiveness studies. Knowing about the costs of an efficacy study will give little valid information relevant to the costs of delivering the treatment in day-to-day practice.

Even when the choice of appropriate outcomes is straightforward, it remains very difficult to decide what constitutes a clinically significant improvement, as opposed to a statistically significant one. These are of course not equivalent – given a large enough sample, it might be possible to show a statistically significant difference between intervention groups, but the difference thus measured might be trivial in clinical terms. Numerous formulae have been proposed to define clinical significance, which give widely differing results (Hafkenscheid, 2000). One simple solution is to reduce the results into dichotomous variables (such as the proportion of patients meeting criteria for remission or the proportion showing 50% reduction in BDI score) and then using these proportions to calculate the Number Needed to Treat (NNT) (Cook & Sackett, 1995). This reflects the number of patients who would need to have the intervention for one patient to show the defined benefit. However, dichotomizing data in this way often involves making arbitrary decisions about the patients' improvement. For example, Bower *et al.* (2003) suggested a 10-point change in BDI score as clinically significant. In contrast, Simpson *et al.* (2003) based their power calculation on the fact that in a sample of individuals with chronic depression even a 3.5-point change in BDI may be clinically important.

Psychological treatment RCTs are difficult and expensive to mount. Furthermore, they are almost entirely dependent on charitable and research foundation rather than commercial funding. In some instances, this means the sample size, number of treatment centres included or the length of post-therapy follow-up may be compromised as compared to drug trials. However, we would argue that a core characteristic of an effective psychological treatment is that the benefit would be expected to endure after therapy has terminated. This is a fundamental difference from a drug treatment, as medication is only effective while the individual continues to adhere to it (Scott, 2000). This has important implications for study design. First, it may reduce the need for 'placebo therapy' that controls for the contact time and non-specific improvements during the acute treatment phase. Secondly, it suggests that the power calculations for a study should be based on relapse rates or survival time at 12 or more months post-therapy rather than focusing only on the immediate outcomes of therapy (Scott, 2001). Demonstrating enduring benefits is likely to favour the intervention as the source of the improvement, rather than non-specific effects, which are not expected to persist beyond the end of the intervention. However, different designs may be required for relapse prevention studies (Paykel *et al.* 1999).

## PROCESS VERSUS OUTCOME RESEARCH

As well as measuring outcomes, an essential component of psychotherapy research is to understand what mediates and moderates these outcomes. Purists argue that explanatory and pragmatic trials should be undertaken separately (Schwartz & Lellouche, 1967). However, all researchers agree that both elements are required if we are to understand the potential role and limitations of a specific psychological therapy. A new drug would not be tested in an efficacy RCT without laboratory research on its mode of action (Parry, 2000). Likewise, we wish to understand the mechanism of action of effective therapies or the common factors that may be associated with good outcome. However, combining process analysis with outcome research is complex, as the nature and size of the sample required might differ markedly, as do the timing of assessments for these two investigations.

A basic requirement in studies of the benefits of a psychological treatment is information on the conduct of the therapy itself. Did the therapists adhere to the therapy model and did they deliver the therapy competently? Quality control is much harder to achieve in psychotherapy RCTs than with drugs but information equivalent to the adequacy of the dose and duration of treatment is relevant. In addition, therapist factors such as their response to issues that emerge in the course of treatment, their allegiance to a particular therapy model and therapist–patient interactions will influence engagement in therapy and treatment outcome (Roth & Fonagy, 1996). These variables need to be addressed or highlighted in the reporting of studies. The obvious way to reduce their effects is to exclude the therapist altogether. The study by Proudfoot *et al.* (2003) did so by delivering a fixed programme of cognitive and behavioural techniques via a computerized treatment package. The study sample, similar to those in many psychological treatment studies (75% female, 88% Caucasian and >66% educated beyond 18 years), appeared to gain significant benefit from the additional intervention. However, even without a therapist other process factors need to be considered. First, as already noted, 35% of those allocated to the computerized treatment dropped out of the study. This is about the same proportion leaving drug or therapy arms in other RCTs. Also, GPs were not blind to the effects of the computer intervention and were given feedback at the end of every session. It would be interesting to know if or how this influenced the GPs' response to the patients allocated to this intervention, particularly as 25% of this group received further treatment during the course of the study (ranging from antidepressants to admission to hospital).

Lastly, it is not always possible to explore an individual's 'readiness' for therapy within the constraints of an RCT as the trial protocol usually reduces or eliminates the behavioural choices of clinicians and subjects. However, researchers have increasingly aimed to explore the factors that dictate whether a treatment is acceptable. As such, the paper by Gabbay *et al.* (2003) makes fascinating reading. The authors highlight that agreeing the agenda is a key therapeutic process in clinical interviews. In primary care, it is the first step in making acceptable a referral to psychological treatment. Gabbay and colleagues explore the degree of agreement about the nature of the individuals presenting problems between participants in an RCT and their general practitioner. Only 33% of patients and general practitioners were in complete agreement about the clients' main presenting problem. Although level of agreement accounted for less of the outcome variance than in previous studies, it emphasizes another difficulty for those treating psychological problems in primary care. In routine practice, the GP may have the option of a series of interviews to explore the patients' explanatory model or map and to negotiate if or when referral may be appropriate (Williams & Healy, 2000). However, when conducting RCTs, patient refusal will lead to early attrition of the study sample. There are no simple answers to this problem although a number of researchers advocate the use of partially randomized preference (Brewin & Bradley, 1989; Parry, 2000).

## CONCLUSIONS

Some researchers argue that results of RCTs in psychotherapy are of little value (Seligman, 1996) and many clinicians are deeply sceptical about using such studies to change clinical practice (Parry, 2000). We think this is nihilistic. All research evolves. Although empirical research in psychological

treatments is in its infancy, it is developing rapidly. The papers published in this issue of the journal are not perfect, but the researchers have grappled with difficult circumstances and sometimes had to make compromises in the methodology employed to ensure that they deliver a study that also has external validity and is meaningful to those working in primary care. Just because psychotherapy RCTs are difficult to design and to interpret, does not mean that they should be abandoned. All the publications in this issue help move our knowledge forward another step. We agree we should not make premature decisions about clinical practice without systematic and reliable data. However, we cannot continue in the paradoxical situation that the most widely available psychotherapies in clinical practice are actually the least researched. Researchers should continue to try to rise to the challenge of overcoming as many of the limitations of RCTs as they can in their study designs. But we also need journal editors to publish such studies. This will increase knowledge of outcomes and contribute to the debate on RCT methodologies and other approaches to outcome research in psychological treatments. Those reading such papers have a responsibility to interpret what they read as they do other published research, with rigorous application of the principles of critical appraisal and evidence-based practice. We applaud the authors of the four high quality studies of a difficult area in this issue.

JAN SCOTT AND TOM SENSKY

## REFERENCES

- Bower, P., Rowland, N. & Hardy, R. (2003). The clinical effectiveness of counselling in primary care: a systematic review and meta-analysis. *Psychological Medicine* **33**, 203–215.
- Brewin, C. & Bradley, C. (1989). Patient preferences and randomized clinical trials. *British Medical Journal* **299**, 313–315.
- Cook, R. J. & Sackett, D. L. (1995). The number needed to treat: a clinically useful measure of treatment effect. *British Medical Journal* **310**, 452–454.
- Gabbay, M., Shiels, C., Bower, P., Sibbald, B., King, M. & Ward, E. (2003). Patient-practitioner agreement: does it matter? *Psychological Medicine* **33**, 241–251.
- Guthrie, E., Moorey, J. & Margison, F. (1999). Cost-effectiveness of brief psychodynamic-interpersonal therapy in high utilizers of psychiatric services. *Archives of General Psychiatry* **56**, 519–526.
- Hafkenschied, A. (2000). Psychometric measures of individual change: an empirical comparison with the Brief Psychiatric Rating Scale (BPRS). *Acta Psychiatrica Scandinavica* **101**, 235–242.
- Keller, M., McCullough, J. & Kelin, D. (2000). The acute treatment of chronic depression: a comparison of nefazadone, cognitive behavioral analysis system psychotherapy, and their combination. *New England Journal of Medicine* **342**, 1462–1430.
- King, M. (1995). Evaluating the benefits of general practice based counseling services. In *Research Foundations for Psychotherapy Practice* (ed. M. Aveline and D. Shapiro), pp. 281–298. John Wiley & Sons: Chichester.
- Mainert, C. (1986). *Clinical Trials: Design, Conduct, and Analysis*. Oxford University Press: Oxford.
- Moher, D., Schulz, F., Altman, G. & Lepage, L. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* **357**, 1191–1194.
- Paykel, E., Scott, J., Teasdale, J., et al. (1999). Prevention of relapse in residual depression by cognitive therapy: a controlled trial. *Archives of General Psychiatry* **56**, 829–835.
- Parry, G. (2000). Evidence based psychotherapy: special case or special pleading? *Evidence Based Mental Health* **3**, 35–36.
- Persons, J. B. & Silberschatz, G. (1998). Are results of randomized controlled trials useful to psychotherapists? *Journal of Consulting Clinical Psychology* **66**, 126–135.
- Proudfoot, J., Goldberg, D., Mann, A., Everitt, B., Marks, I. & Gray, J. A. (2003). Computerised, interactive, multimedia cognitive behavioural therapy for anxiety and depression in general practice. *Psychological Medicine* **33**, 217–227.
- Roth, A. & Fonagy, P. (1996). *What Works for Whom?* Guilford Press: New York.
- Scott, J. (2000). Treatment of chronic depression. *New England Journal of Medicine* **342**, 1518–1520.
- Scott, J. (2001). Cognitive therapy of depression. *British Medical Bulletin* **57**, 101–113.
- Seligman, M. E. (1996). Science as an ally of practice. *American Psychology* **51**, 1072–1079.
- Schwartz, D. & Lellouch, J. (1967). Explanatory and pragmatic attitudes in therapeutic trials. *Journal of Chronic Disease* **20**, 637–648.
- Simpson, S., Corney, R., Fitzgerald, P. & Beecham, J. (2003). A randomized controlled trial to evaluate the effectiveness and cost-effectiveness of psychodynamic counselling for general practice patients with chronic depression. *Psychological Medicine* **33**, 229–239.
- Smith, M., Glass, G. & Miller, T. (1980). *The Benefits of Psychotherapy*. Johns Hopkins University Press: Baltimore.
- Streiner, D. L. (2000a). The 2 'Es' of research: efficacy and effectiveness trials. *Canadian Journal of Psychiatry* **47**, 552–556.
- Streiner, D. L. (2002b). The case of the missing data: methods of dealing with dropouts and other research vagaries. *Canadian Journal of Psychiatry* **47**, 68–75.
- Williams, B. & Healy, D. (2001). Perceptions of illness causation among new referrals to a CMHT: explanatory model or explanatory map? *Social Science & Medicine* **53**, 465–476.