Mental health services research methodology

MIKE SLADE¹, ELIZABETH KUIPERS¹ & STEFAN PRIEBE²

¹Institute of Psychiatry, London & ²Bart’s and The London School of Medicine, London, UK

Summary

Evidence-based mental health is an important goal, and randomized controlled trials (RCTs) are currently used as the currency. Significant gains have been made in overcoming technical difficulties with RCTs, but conceptual issues with the use of RCTs as ‘best’ evidence can also be identified. Some limits of RCTs for research into individual patients, local services, and national policy will be identified. The central thesis is that RCTs have an important contribution to make, but are only one form of evidence. Another framework for research—realistic evaluation—is described, in which the context and mechanisms of action are considered, as well as the outcome. Realistic evaluation will lead to different forms of evidence, including but not limited to RCTs, and will be more illuminating for some research questions than solely consider RCTs.

Introduction

Biomedical science methods are widely used within mental health services research. These methods involve the empirical testing of hypotheses deduced from theory, and involve research methodologies which allow replication and generalization. The methodological standard bearer is the randomized controlled trial (RCT), since a control group allows the inference of change due to the treatment, rather than other causes, and randomization minimizes bias at the point of inclusion. The importance of biomedical science methodologies has led to the proposal that a hierarchy of evidence exists, with systematic reviews and meta-analyses of RCTs at the top, followed by RCTs with definitive results, RCTs with non-definitive results, cohort studies, case-control studies, cross-sectional surveys and case reports (e.g. Greenhalgh, 1997). The ‘unit of currency’ for this hierarchy is the RCT, with pooled groups of RCTs being the best form of evidence, followed by individual RCTs, in turn followed by studies which do not possess all the characteristics of RCTs. The implication of this is that RCTs are the ‘gold standard’ of evidence.

This focus on RCTs brings many benefits: uncontrolled or poorly controlled studies are given less weight, more importance is attached to methodological issues, and more caution is exercised in evaluating outcome. However, another result is that research questions are designed so that they can be answered by RCTs. Specifically, the use of RCTs involves the identification of an intervention which is given to the patients in the experimental group, but not in the control group. This encourages the asking of particular types of research questions, typically of the form ‘Does intervention X work for disorder Y?’ It will be argued that the RCT methodology limits the questions that can be asked, and hence can restrict the potential findings from research. Furthermore, if different questions were being asked, then RCTs would not always be the best methodology to employ.

Current research methods

The methods of the biomedical sciences have many strengths. The emphasis on empiricism has given rise to a culture of statistical and methodological rigour. The ability to characterize the strength of a finding has indicated where further work is needed. The development of meta-analytic technologies has maximized the use of available data. In particular, statistics such as ‘Number Needed to Treat’ allow the common patient question of ‘How likely is it that this will help me?’ to be answered, with a probabilistic statement.

As with any methodology, a number of technical problems with RCTs have also been identified. For example, the meaning of available evidence is not always understood, with the danger that ‘unproved’ is equated with ‘disproved’, which may account for the emphasis on medication over psychosocial interventions in schizophrenia (e.g. Gilbody & House, 1999). Other identified issues include recruitment bias, differential expectations, problems with randomization and obtaining informed consent.
maintaining blindness; insufficient follow-up, methodological bias (e.g. insufficient use of intention-to-treat analyses), and defining 'improvement'. However, creative and methodologically-sound solutions to all these problems have been proposed (Blacker & Mortimore, 1996; Bradley, 1997; Taylor & Thornicroft, 1996). Technical problems are all potentially amenable to methodological resolution.

Conceptual problems are not so easily resolved. The biomedical sciences methods have been developed most fully in drug trials, in which both the object of enquiry (the dose of medication) and the entity in which change will be measured (the patient) are relatively clear. Conceptual limitations become most evident when considering attempts to use these methods in other domains. The limitations will be illustrated with reference to the three levels proposed in the matrix model for mental health services: patient, locality, and regional/national (Thornicroft & Tansella, 1999).

Conceptual problems at the patient level

Implicit in the RCT methodology is the assumption that the intervention can be given in the same way and to the same extent to each person in the experimental group, and that its mechanisms of action are the same for whomsoever receives it. In other words, it assumes that there are no differences in any relevant respect in either the multiple instances of the intervention (providing treatment fidelity is ensured) or between the people in the experimental or control groups (providing allocation is random and hence any differences are due solely to chance). The most widely used application of this method has been in medication, resulting in a large and expanding evidence base for pharmacotherapy. However, the natural science methods have also been applied to other interventions, such as psychotherapy. Because of the methodology to be employed, the question that has been investigated is 'Does intervention X work for condition Y?' The problem is then how to make what is an essentially interpersonal intervention sufficiently uniform that it can be provided in the same way to each person. The solution adopted has been to utilize manualized protocols to ensure treatment fidelity, supported by the development of assessments to measure the extent to which the treatment accords with the protocol. This has resulted in clear findings of high relevance to mental health services (e.g. Kuipers et al., 1997; Sensky et al., 2000).

However, there is a problem with this approach. Some therapists are better than others (Roth & Fonagy, 1996). This difference can be exhibited in a number of domains of competence (Bryant et al., 1999). This is not solved by identifying skill levels, and differentiating between expert therapists, competent therapists and so on, since there will be an interaction between patient and therapist characteristics—some patients will 'connect' with some therapists, and some with others. Since therapists are also people, differences apply not just to skills and techniques, but also to personal characteristics such as appearance, sense of humour, and speech accent. Many aspects may impact on the outcome of therapy. Given the complexity of a dynamic process of interaction, attempting to control fully for individual differences through larger, more targeted studies may prove an elusive goal.

RCTs undoubtedly have an important role to play in answering the question 'Does intervention X work for condition Y?' However, the question 'Which patients with condition Y does intervention X work for?' may prove to have more clinical relevance, and answering this question may involve asking the question 'How does intervention X work?', a question which cannot be answered just by using RCTs.

Conceptual problems at the local level

The difficulties with biomedical sciences methods become more pronounced with investigation into service-level interventions. The research question typically asked at this level is 'Does service configuration X work with group Y?' This has involved identifying comparable experimental and control programmes (e.g. contiguous catchment areas), ensuring fidelity of programme administration, and before and after measurement of change in a broad range of areas. Recent examples include the UK700 study of case management (Burns et al., 1999b) and the PRiSM Psychosis Study of community care (Thornicroft et al., 1998).

The UK700 study has also been criticized, for failing to differentiate between case management and assertive community treatment models of care (McGovern & Owen, 1999). The response by the authors has been to highlight the similarities between the two models (Burns et al., 1999a). Disagreement therefore focuses on what intervention was being evaluated. Although not an RCT, the PRiSM Psychosis Study also illustrates the difficulty of service-level research using the RCT approach of giving an intervention to a group of patients, where the intervention is a service configuration. The study has been criticized for employing a mixed model of care (Marshall et al., 1999), and for inadequately specifying the structures and processes of the interventions (Sashidharan et al., 1999). The response by the authors to these criticisms has been to highlight that the study was intended to test differently configured catchment-area services, rather than different models of care (Thornicroft et al., 1999). Disagreement therefore again seems to revolve around what intervention was being evaluated.

This line of critique and defence is predictable, since the intervention is a social programme, and is
In different examples of the 'same' programme, there will be important differences in resources (such as quality of buildings, locations relative to patients, amount of money for continuing professional development of staff), processes (what is done, by whom, and in what way), and structures, such as the debate regarding the necessary components of assertive community treatment (e.g. Deci et al., 1995; McGrew & Bond 1995; Teague et al., 1998). Indeed, it would be easy to compile a list of several hundred service characteristics which may impact on outcome for specific patients. Most of these factors are not measured, especially given the lack of standardization regarding what characteristics of a service to report. The limited efforts to develop standardized assessments of services, such as the European Service Mapping Schedule (Johnson et al., 1998), have highlighted the complexity of characterizing services. Therefore, although in any individual study random allocation will ensure that any initial differences in the two groups are due to chance, it will be impossible to generalize the findings, since the 'intervention' (the social programme) will be inadequately characterized (Slade & Priebe, 2001).

High-quality RCTs have an important contribution to make in generating evidence about programmes of care. Perhaps the best hope for the RCT approach, as expressed by the UK700 authors, is that 'Real progress will be made when essential ingredients in complex interventions are individually subject to equally rigorous evaluation' (Burns et al., 1999a, p. 1386). However, even such a knowledge base would be unable to account for interactions between different components of the programme, or emergent properties of the system. The goal of ensuring internal validity by adequately characterizing all relevant characteristics whilst retaining external validity such that what is investigated generalizes to what can be done in routine practice may be impossible to meet. RCTs cannot generate all the necessary evidence for developing mental health services.

**Conceptual problems at the national level**

It is at the national level that the conceptual problems with biomedical sciences methods become most prominent. Questions which are asked at this level include: How will a higher national expenditure on mental healthcare impact on outcome criteria? What are the effects of different approaches to distributing this money? What is the best balance between central and localized control of spending? What is the best balance between clinical governance and professional self-regulation? What effects can be expected from the introduction of Community Treatment Orders? How can stigma be reduced? What type of mental health research funding should be prioritized?

For all of these questions, the type of information which is used in practice to make decisions does not routinely come from RCTs. For some questions, current practice is based on anecdote—community treatment orders worked elsewhere, so they should work here. For others, current practice is based on precedent—current resource allocation formulae have been developed iteratively from previous estimates (Jarman et al., 1992). Current practice is not based on evidence of the RCT form (Kane, 2002). Indeed, consider what gathering such evidence would involve. To decide how much to spend on mental healthcare in England, for example, one could design a study which used Health Authorities as cases. But to identify the effectiveness of the intervention (funding level) a number of factors need to be controlled for, including characteristics of the population, current service development levels, current levels of mental health spending, and population-based needs assessment. One would soon run out of Health Authorities for matching. Similarly, what would be the intervention? Give half the Health Authorities high funding and half low, and investigate the resulting health gain? But all Health Authorities would want more money, so some would feel they are getting a bad deal from this trial, leading to a demoralized work-force, whose more able members might move to working in nearby Health Authorities who were randomized into high funding. What if some Health Authorities decided to invest heavily in primary care services, some in specialist services such as early intervention in psychosis, and some in generic mental health services? This would immediately confound the trial, since health gain differences might be due to service configuration rather than money spent.

**Unresolvable problems**

The above example is elaborated to underline the fact that some questions simply cannot be investigated using RCT methodologies, and for other questions RCTs are not the best form of enquiry. For example, RCTs require the use of groups of patients who differ only by chance in all relevant respects. The technical solution is large RCTs—'mega-trials'—to allow the 'true' effect size to be identified. However, it may not be practical to undertake RCTs of a size sufficient to discriminate between groups when investigating complex psychological, interpersonal, social, ethnic, cultural, ethical or political questions. Examples of research questions which are difficult to address using RCTs are shown in Table 1.

Some of these areas are, of course, subject to substantial research efforts using RCTs, by changing the question to fit the methodology. For example, by changing the individual differences question to
Table 1. Examples of research questions which are difficult to investigate using RCTs

<table>
<thead>
<tr>
<th>Level</th>
<th>Topic</th>
<th>Research question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td>Culture</td>
<td>Is it better to be seen by a therapist of the same cultural background?</td>
</tr>
<tr>
<td></td>
<td>Ethnicity</td>
<td>Do service-related factors account for any of the association between being Afro-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Caribbean and compulsory admission?</td>
</tr>
<tr>
<td></td>
<td>Social context</td>
<td>Should a patient whose depression occurs in the context of domestic violence be</td>
</tr>
<tr>
<td></td>
<td>Individual differences</td>
<td>prescribed anti-depressants?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Will this treatment work for this patient?</td>
</tr>
<tr>
<td>Local</td>
<td>Social programmes</td>
<td>What skills and competencies are needed in this team?</td>
</tr>
<tr>
<td></td>
<td>Inter-agency working</td>
<td>How can communication be improved between health and social services?</td>
</tr>
<tr>
<td></td>
<td>Mixed economy of care</td>
<td>What is the best balance between voluntary and statutory sector provision of services?</td>
</tr>
<tr>
<td></td>
<td>Service structures</td>
<td>Should we start an assertive community treatment team in this area?</td>
</tr>
<tr>
<td>National</td>
<td>Social change</td>
<td>How do we get the media to report mental illness in a more balanced way?</td>
</tr>
<tr>
<td></td>
<td>Relationship with other funding</td>
<td>Which Government department should be responsible for long-term nursing care?</td>
</tr>
<tr>
<td></td>
<td>demands</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Role of professions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Research funding priorities</td>
<td></td>
</tr>
</tbody>
</table>

'Does this treatment work for groups of patients similar in some defined way to this patient?' allows the generation of Number Needed to treat data. However, the findings are often equivocal, for example sometimes yes, sometimes no' (for the culture question) or 'it helps some patients, so it's worth trying' (for the social context question). The technical solution of matching control and experimental groups on all relevant characteristics cannot sufficiently account for the differences which impact on the intervention, leading to inconsistent and non-generalizable results. Although these problems occur in pharmacotherapy research, they become more evident when evaluating psychotherapy, service configurations and national programmes, due to the increasing complexity of the intervention and the difficulty in operationalizing all defining elements.

The development of knowledge in the physical sciences happens through a process of progressive development, testing and refinement of causal hypotheses. Refinement focuses, at least partly, on the limits within which proposed rules appear to hold good, and the imperfections in the hypotheses giving rise to these limits. The limits are often as illuminating as the rules. This article has highlighted some of the limits, and it will now be argued that what is needed as well is a different type of research, aimed at answering different questions. Such a framework may be offered by an alternative methodology, which has been termed 'Realistic Evaluation' in a recent book (Pawson & Tilley, 1997).

Another framework for research: realistic evaluation

Realistic evaluation is the process of evaluating the effectiveness of particular social programmes targeted at specific social problems. Pawson & Tilley (1997) discuss social programmes, which they define as 'the interplays of individual and institution, of agency and structure, and of micro and macro social processes' (p. 63). The central message of the book is the need to move from a successionist model to a generative model of causation (Harré, 1972). Successionist theory holds that causation is unobservable (following Hume, 1739), and observational data are the only mechanism for inferring causality. This theory leads directly to the methods of experimental manipulation, and the post-post-comparison of experimental and control groups which permeate mental health services research today. In other words, an observed statistical association between a defined service configuration and individual outcomes is the basis for predicting effects of services, since a full understanding of why a service achieves a more or less favourable outcome is not possible. The dynamic processes linking service configuration and
outcome remain unknown. Generative theory, by contrast, holds that there is an observable connection between causally connected events, and that internal features of the thing being changed are central to the understanding of causality. To illustrate, a successionist notion of causality is involved in the statement 'gravity causes an apple to fall to Earth', and a generative understanding in the statement 'bereavement causes depression'. Generative theory suggests that 'causal outcomes follow from mechanisms acting in contexts' (Pawson & Tilley, 1997, p. 58). An understanding of the causal mechanisms linking input with outcome and of the contextual factors influencing these processes provide the basis for a prediction of what may happen in a concrete situation.

So what is the relevance of this to mental health services research? The above definition of a social programme accords exactly with the subject matter of mental healthcare, whether the 'intervention' be a psychiatrist prescribing for a patient, the development of an assertive community treatment service, or attempts to reduce the stigma of mental illness. It is proposed that the current ethos of 'evidence-based mental health', where the term 'evidence' is equated with RCTs and meta-analysis (i.e. successionist methods), can only provide one form of evidence. The seminal example of this fate in other areas of social research is the review by Martinson (1974) of offender rehabilitation programmes. The review considered all published reports in English between 1945 and 1967, and the full version ran to 1400 pages. He concluded:

I am bound to say that these data, involving over two hundred studies and hundreds of thousands of individuals as they do, are the best available and give us very little reason for hope that we have in fact found a sure way of reducing recidivism through rehabilitation. This is not to say that we have found no instances of success or partial success; it is only to say that these instances have been isolated, producing no clear pattern to indicate the efficacy of any particular method of treatment. (Martinson, 1979, p. 49)

As Pawson & Tilley (1997) note, the problem at one level is the impossible criteria for success, in which an intervention 'works' only if it produces positive outcomes in all trials in all contexts. The RCT proponent might argue that the Martinson review simply did not have access to an adequate (i.e. RCT-based) evidence base. However, the pattern of developments in mental health research is depressingly similar, with an increased emphasis on methodologies which cost more and more to implement (larger samples, increased programme fidelity, etc.), in the belief that interventions will ultimately be categorized into 'effective' and 'ineffective'.

Pawson & Tilley (1997) use as an example the installation of closed circuit television cameras in an attempt to reduce thefts in car parks. The authors identify eight possible mechanisms by which the intervention could reduce crime and six contextual issues which could limit the potential for some of these effects. Even if only even a few of these were really operative, it is predictable that testing the intervention using RCTs will produce conflicting results. The best hope of progress is to develop and test a series of more detailed hypotheses, based on causal models of how the approach might be operating. The relevance to mental health services research is the challenge of whether the right questions are being asked—or rather being asked in the right way.

The overall approach of realistic evaluation is to go back a stage in scientific research. It involves starting with hypotheses about mechanisms which produce particular outcomes, and the context within which these mechanisms operate. These hypotheses suggest initial patterns likely to be found in whatever problem is being investigated. When a specific intervention is targeted at altering particular aspects of the mechanism in particular ways, hypotheses about the relationship between context and outcome allow the formulation of specific research questions. As in meta-analysis, studies seldom stand alone. However, the goal of grouping studies is to identify 'middle range theories', which lie between minor hypotheses and all-inclusive systematic theories of social programmes (Merton, 1968). The central difference between this approach and the current focus on RCT methodology is that theories of mechanism under test and the contextual issues hypothesized to be influencing them need to be spelt out, so that related studies (including but not limited to RCTs) can be linked. This already happens to some extent, notably in pharmacotherapeutic and psychological intervention research. However, middle range theories do not appear to be central to research at the local and national levels.

Implications for mental health services research

What kind of studies are undertaken within this research framework? There would be more focus on identifying mechanisms of change and contexts in which these mechanisms are activated. One approach is to use RCTs, as is beginning to be apparent in trials of psychological interventions. For example, the London and East Anglia RCT of cognitive-behavioural therapy for psychosis identified 'response to hypothetical contradiction' as a moderator of outcome (Garety et al., 1997). This both accords with the cognitive model of schizophrenia, and is of practical clinical utility when assessing an individual for suitability for treatment. However, investigating contexts and mechanisms involves more detailed questions than RCTs have been designed to answer, indicating a need for a broader range of methodologies.
Table 2. Possible context-mechanism-outcome combinations for an early psychosis service

<table>
<thead>
<tr>
<th>Domain of improved outcome</th>
<th>Mechanism</th>
<th>Context (compared with standard service)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remoralization</td>
<td>Improved self-esteem through being seen by a specialist service</td>
<td>Service is described as 'specialist' by referrer</td>
</tr>
<tr>
<td></td>
<td>Contact with motivated staff</td>
<td>High levels of within-team support</td>
</tr>
<tr>
<td></td>
<td>Being seen outside of mental health services</td>
<td>Service is not seen as part of mental health services</td>
</tr>
<tr>
<td>Remediation</td>
<td>Iatrogenic effects are minimized</td>
<td>Service is less institutional than standard care</td>
</tr>
<tr>
<td></td>
<td>Specific intervention is of benefit to the individual patient</td>
<td>Service provides relevant intervention with expertise</td>
</tr>
<tr>
<td></td>
<td>More attention from expert staff</td>
<td>Low caseloads, high expertise in team</td>
</tr>
<tr>
<td>Rehabilitation</td>
<td>Improved coping strategies</td>
<td>Team uses stress/vulnerability model of psychosis, rather than biological model</td>
</tr>
<tr>
<td></td>
<td>Relapse warning signs identified sooner</td>
<td>Early warning signs work done with more patients</td>
</tr>
<tr>
<td></td>
<td>Long-term engagement</td>
<td>Patient gets on with team members</td>
</tr>
</tbody>
</table>

At the local level, consider an attempt to evaluate a service aspiring to offer early intervention in psychosis. Possible mechanisms for improved outcome (compared with standard care) in terms of remoralization, remediation and rehabilitation are shown in Table 2.

Each context-mechanism combination indicates a research question. Since multitudinous mechanisms might account for improved outcome, attempts to evaluate an early psychosis service as identical in all implementations will result in conflicting findings. What is needed is to identify what putative mechanism is being investigated, and what context is required for that mechanism to be activated. Once identified, a range of methodologies might then be appropriate for testing hypotheses, including but not limited to RCTs.

At the national level, research needs to draw from other disciplines (e.g. marketing, politics), in a bid to understand population-level mechanisms of change. For example, the stigma of homosexuality has probably been reduced in a number of ways, probably including legislative changes, reductions in censorship, the dissemination of research findings, and vociferous lobbying. Identifying the mechanisms of action for these various strategies may have implications for reducing the stigma of mental illness.

The result of this research agenda would be an understanding, rather than an explanation, of what aspects of a social programme produce change in what way, and for what people. Only in this way will mental health services move towards being rationally planned, developed and evaluated. RCTs will still have an important role, but the term 'evidence' will have a broader meaning, encompassing the results of many types of research which investigate many types of questions. Meaningful and important evidence is likely to come, for instance, from individual case studies (Alderman, 2002), anthropological research (Bartlett, 2002), and qualitative studies (Williams, 2002). If the ultimate goal of the mental health system is to operate on the basis of realistic evidence, rather than historical precedent or clinical anecdote, then the challenge for health planners, practitioners and researchers is to understand not only which services work, but also why, how, when and where. New methodologies have to be developed and applied in mental health services research for achieving such an understanding.

Acknowledgements

The ideas for this article have been developed through discussion with many colleagues, including Jonathan Bindman, Derek Bolton, Gyles Glover, Morven Leese, James Tighe and Graham Thornicroft.

References


