Improving the delivery and organisation of mental health services: beyond the conventional randomised controlled trial

SIMON GILBODY and PAULA WHITTY

Background There is an ethical imperative to evaluate service and policy initiatives, such as those highlighted in the recent National Service Framework, just as there is to evaluate individual treatments.

Aims To outline the best methods available for evaluating the delivery and organisation of mental health services.

Method We present a narrative methodological overview, using salient examples from mental health services research.

Results Cluster randomised studies involve the random allocation of groups of clinicians, clinical teams or hospitals rather than individual patients, and produce the least biased evaluation of mental health policy, organisation or service delivery. Where randomisation is impossible or impractical (often when services or policies are already implemented), then quasi-experimental designs can be used. Such designs have both strengths and many potential flaws.

Conclusions The gold standard remains the randomised trial, but with due consideration to the unit of randomisation. Use of quasi-experimental designs can be justified in certain circumstances but should be attempted and interpreted with caution.

Declaration of interest S.G. is supported by the Medical Research Council Health Services Research Fellowship Programme.

Research in mental health should address questions relating to the best methods of organising and delivering services, and not just questions relating to drug treatments or psychological therapies. This issue is particularly important at the moment for two reasons. First, there are major gaps in the evidential basis of core policy recommendations such as those contained in the Mental Health National Service Framework (NSF) (Secretary of State for Health, 1999; Tyrer, 1999). Second, the National Health Service (NHS) Research and Development programme has been restructured recently into two main streams of funding: Health Technology Assessment (HTA) and Service Delivery and Organisation (SDO) (Department of Health, 2000). The SDO has carried out a 'listening exercise' to identify NHS priorities for its activity, and the Mental Health NSF was one of the areas identified (National Co-ordinating Centre for NHS Service Delivery and Organisation Research and Development, 2000).

METHODS FOR EVALUATING MENTAL HEALTH SERVICE DELIVERY AND ORGANISATION

This article aims to review, using salient examples, the most robust randomised and quasi-experimental study designs available for evaluating the delivery and organisation of mental health services. The principal study designs covered here are:

(a) cluster randomised controlled trials (RCTs);

(b) 'controlled before and after' (CBA) studies;

(c) interrupted time series (ITS) studies.

Given the particular importance of the cluster RCT (Ukoumunne *et al*, 1999*b*), we cover the methodological issues involved in this design in the most detail.

Cluster randomised controlled trials

Randomisation has rightly remained the most robust method of eliminating selection bias and judging the true value of interventions in mental health (World Health Organization, 1991), as in all health care (Sackett *et al*, 1991). For questions relating to the best method of delivering and organising mental health services or about the best method of changing clinical practice for the better (Wells, 1999), randomisation still may be appropriate but it becomes either impossible or inappropriate to randomise individuals (Campbell *et al*, 2000*a*). Instead it is individual hospitals, geographical areas, clinical teams or groups of clinicians and, consequently, all the patients they treat, that should be randomised. When 'clusters' of individuals are randomised in this way, there are profound implications for the design and analysis of studies that are only just beginning to be realised (Campbell & Grimshaw, 1998).

There are sound theoretical, ethical and practical reasons why clustering should occur in mental health services research, which the following examples serve to illustrate. First, innovative methods of delivering a service often can be offered only at the level of the hospital or general practice (Buck & Donner, 1982; Gail *et al*, 1996). For example, if a new specialist outreach service for the management of deliberate self-harm were to be instituted in a hospital where the previous policy was one of usual care and non-compliance with national guidelines (Hawton & James, 1995), then it would be difficult to ask individual clinicians to randomise their individual patients to receive this service. However, from a policy perspective, given alternative models of intervention for this group (Hawton *et al*, 1998), it is important to judge the true clinical and cost effectiveness of this policy. The most rigorous way to establish this is to randomise individual hospitals to introduce this service or to continue with their current model of service delivery. Similarly, efforts to examine the true value of counsellors in general practice have shown that it can be difficult to get general practitioners (GPs) to randomise their patients to receive counselling or usual care, resulting in recruitment problems to apparently well-planned trials (Fairhurst & Dorwick, 1996). A practical solution is to randomise individual GPs or practices to receive a counsellor if studies are to be feasible and sufficiently powered.

https://doi.org/10.1192/bjp.180.1.13 Published online by Cambridge University Press

A second good reason to randomise clinicians rather than their individual patients is the problem of 'contamination' between subjects (Buck & Donner, 1982). This is especially problematic when evaluating behavioural interventions designed to change clinical practice, such as practice guidelines (NHS Centre for Reviews and Dissemination, 1994). For example, when evaluating interventions designed to encourage GPs to recognise and appropriately manage depression, it is difficult for them to switch on this skill for one patient and then resume usual practice for the next patient. Hence, individual GPs (or practices) must be randomised to give this intervention.

What are the implications of randomisation by cluster?

The main implication is the larger sample size required for cluster trials. It should be self-evident that people in clusters will share similarities. For example, people in one geographical hospital catchment area will be more similar to each other than they are to other people in another area, by reason of geographical variations in case mix measures, such as affluence, deprivation, age and ethnic mix (Rice & Leyland, 1996; Smith *et al*, 1996). Equally, general practices in the same geographical area may attract different patient profiles by virtue of the ethnicity, gender or ethos of their GPs. Such socio-demographic and practice style variables are potentially related to health need and outcomes of treatment for these clusters of patients (Rice & Leyland, 1996).

Conventional statistical methods assume that individual observations and responses are, in fact, independent of each other. Consequently, when clustering occurs and there is an intracluster correlation of variables associated with outcome, then calculations such as sample size, assessments of power and judgements of statistical significance cease to be valid (Kerry & Bland, 1998*b*,*c*). When intracluster correlation is present and this is ignored by analysing at the individual patient level, then necessary sample size is underestimated and significance is overestimated, making type I and type II errors more likely. This is often called a 'unit-of-analysis error' (Whiting-O'Keefe *et al*, 1984; Divine *et al*, 1992).

How to cope with clustering

The key advantage of randomised studies is that group allocation occurs according to the play of chance and both known and unknown confounding variables in the long run will be evenly distributed, thus minimising baseline differences. However, when there are relatively few clusters such as individual hospitals or practices to be randomised, then the chances of baseline imbalances arising by chance alone increase dramatically. This can be anticipated by the use of matched pairs, whereby clusters are matched according to known variables associated with outcome and one cluster from each pair is allocated to a group according to a random process (Meinert, 1986).

The effect of clustering and correlation of outcomes also needs to be accounted for in the design and analysis of both randomised and non-randomised studies (Buck & Donner, 1982), although researchers rarely appreciate this (Donner *et al*, 1990; Simpson *et al*, 1995; Campbell *et al*, 2000*b*). Randomised trials in mental health (both cluster and non-cluster) thus far have been generally underpowered and have shown poor consideration of sample size requirements (Hotopf *et al*, 1997; Thornley & Adams, 1998). When estimating necessary sample sizes and analysing the results of cluster trials, then two components must be considered – the level of variation within clusters and the level of variation between clusters. These factors are combined in an important statistic known as the intraclass correlation coefficient (often referred to as the ICC or ρ) (Kerry & Bland, 1998*a*), which is used to inflate the sample size estimate required to achieve a statistical power and to reduce the precision of confidence intervals obtained by traditional statistical analysis (Donner & Klar, 1994). The Appendix shows the statistical considerations used in calculating and inflating sample size for clustered studies.

The value of ρ is between 0 and 1 and its potential influence can be appreciated by the following two hypothetical examples. When $\rho=0$ there is no intraclass correlation in outcomes and individuals can be treated as independent, traditional sample size calculations are valid and no inflation of sample size is required. However, when there is perfect intraclass correlation of outcomes and $\rho=1$, then each cluster must be considered as an independent unit of analysis. Hence, there will need to be as many clusters as there would need to be individual patients stipulated from traditional power calculations. In practice, the minimum sample size (in terms of total numbers of individual patients) required to produce a sufficiently powered cluster study tends to be about twice that required for a non-cluster trial (Campbell & Grimshaw, 1998).

Unfortunately, there is a poor level of knowledge of the importance of intraclass correlation and it is rarely accounted for in design or analysis (Donner *et al*, 1990). Similarly, sample size estimates require a prior knowledge of ρ, and this can be obtained only from research publications (Kerry & Bland, 1998*c*). A survey of major biomedical publications shows that this is rarely reported by authors of any type of study, including those relating to mental health (Donner *et al*, 1990; Ukoumunne *et al*, 1999*b*). Calls therefore have been made to remedy this situation (Elbourne, 1997) by amendments to the CONSORT statement on the proper reporting of clinical trials (Begg *et al*, 1996). Additionally, a database of ICC values recently has been initiated (Ukoumunne *et al*, 1999*b*).

Cluster randomised studies share more similarities than differences with individualised RCTs. For example, the problems of biased assessment of outcome are the same in clustered studies as in individualised studies, and blinded assessment of outcome at the level of the individual patient by those unconnected with the delivery of the service under evaluation always should be attempted when this is required. However, some service-based interventions tend to focus upon administrative outcomes (such as readmission or length of stay), and in these cases blinding is often neither necessary nor possible.

Two examples

The importance of clustering and the challenges that this poses are best illustrated by considering two examples from the psychiatric literature.

Routine screening for emotional disorders by GPs. The knowledge that GPs frequently fail to recognise common disorders such as depression has led some to suggest that well-evaluated screening and outcome questionnaires such as the General Health Questionnaire (GHQ) should be administered to primary care patients and their results fed back to clinicians (Johnstone & Goldberg, 1976). Although the psychometric properties of the GHQ are well known (Goldberg & Williams, 1988), the degree to which the information provided by this

https://doi.org/10.1192/bjp.180.1.13 Published online by Cambridge University Press

test actually changes clinician behaviour is less clear (Rand *et al*, 1988). On the face of it, a simple patient-completed test would seem a useful addition to clinical management.

A potentially informative study is therefore that by Hoeper *et al* (1984) of 2000 US primary care attenders who completed the GHQ in the waiting room. Half of these patients were randomised to have their GHQ score fed back to their GP, whereas the other half had their GHQ scores withheld. Individual patients were the unit of randomisation, and individual GPs saw both patients with GHQ scores and patients without. The authors found that GHQ feedback had minimal effect on the clinical recognition of depression, and concluded that the strategy was of little benefit as a policy (Hoeper *et al*, 1984). However, feedback for some patients may have influenced the management of depression for all patients. The negative outcome of this study therefore is difficult to interpret, given the potential cross-contamination that might have occurred (Gilbody *et al*, 2001). A more robust design would have randomised GPs and included clustering effects in the design and analysis.

Evaluating routine guideline implementation strategies. Practice guidelines and educational interventions provide an alternative strategy in improving the management of depression in primary care (Katon *et al*, 1995). The Hampshire Depression Project (Thompson *et al*, 2000) evaluated this approach among UK GPs. The authors randomised practices, stratifying them according to important variables that determine mental health practice, including fundholding status and resources available to treat mental illness.

An important outcome was an improvement in patients' depression symptomatology. Conventional power calculations indicated that to detect an improvement at follow-up from 40% in the control group to 50% in the intervention group, 400 patients would be required to achieve 80% power at 5% significance. This conventional power calculation takes no account of clustering, and to adjust this they took a best estimate of the effect of clustering from a primary-care-based screening study of cardiovascular risk factors (Family Heart Study Group, 1994). A power calculation incorporating these data inflated the necessary sample size by a factor of 1.48 up to 590 patients.

In analysing patient improvement data, the authors found there to be a very low level of intraclass correlation that required no adjustment. However, another important trial end-point was the clinicians' sensitivity and specificity with which they detected depression, which was correlated within clusters. This required adjustment to avoid overoptimistic estimates of statistical significance or unduly narrow confidence intervals. The authors found no benefit for educational interventions to implement guidelines, and used a robust design and appropriate analysis to reach this conclusion. Further, by publishing their ICCs, they have made the data available for others to calculate sample sizes in future studies.

Quasi-experimental designs

There are also situations where RCTs (individualised or cluster) are not appropriate for answering questions about the organisation and delivery of health services (Black, 1996). For example, a new service may already have been implemented or there may be too few organisational units available to randomise (Jacobs *et al*, 1986). In these cases a robust non-randomised (Black *et al*, 1998) or 'quasi-experimental' evaluation design still can produce valid results (Campbell & Stanley, 1966; Cook & Campbell, 1979). However, researchers and decision-makers need to be aware of their strengths and weaknesses.

Quasi-experiments evaluate the effects of an intervention without using randomisation to create the comparisons from which the effects are inferred (Ukoumunne *et al*, 1999*a*). Two basic types of quasi-experimental study design usually are distinguished: 'non-equivalent group' and 'interrupted time series' designs (Cook & Campbell, 1979). The best-known non-equivalent group design with (potentially) reasonable validity is the "treated group untreated group pre-test post-test design" (Cook & Campbell, 1979), better known as the CBA study.

'Controlled before and after' studies

These are studies in which the responses of an intervention group and a non-randomised comparison (control) group are measured before and after an intervention (Cook & Campbell, 1979). The CBA design is probably the best alternative to a clustered RCT for evaluating an organisational intervention when control groups are available but randomisation is either impossible or inappropriate – although some would argue that where a CBA study is undertaken a clustered RCT probably would have been possible (Campbell *et al*, 1999). The control group needs to be chosen carefully to ensure comparability with the study group at baseline (NHS Centre for Reviews and Dissemination, 1994) and that similar trends and changes are likely to be experienced throughout the duration of the study (Campbell *et al*, 1999). Clearly, there is potential for bias and confounding (an alternative explanation for the outcome) in the estimate of effectiveness from such study designs owing to the non-randomised control group (Cook & Campbell, 1979). However, the key with all non-randomised designs is to try to anticipate all the possible threats to internal validity, and either select the control group accordingly and/or measure the factors concerned in both the intervention and control groups, controlling for these in the analysis. Stratification according to important prognostic factors also can reduce bias in this type of study (Ukoumunne *et al*, 1999*b*).

There remains an ongoing debate about the relative merits of randomised and non-randomised studies in producing unbiased estimates of effect, and a recent systematic review has demonstrated that non-randomised study designs can produce similar results (McKee *et al*, 1998; Reeves *et al*, 1998), provided that the studies are well-designed, executed and analysed. However, this assertion has yet to be examined in mental health. Also, many of the advantages argued for observational studies in general (including the potential for greater external validity because more 'real-world' scenarios can be accommodated in the study design; Black, 1996) are equally applicable to CBA studies.

Many of the organisational interventions in the mental health NSF, currently being implemented without an evidence base, could be evaluated with a CBA design. For example, the staggered introduction of crisis intervention services would be an ideal opportunity if suitable control groups can be identified among the 'late implementers'. However, a recent example of a CBA design within the UK mental health field – the PRiSM Psychosis Study of the Role of Community Mental Health Teams (Thornicroft *et al*, 1998) – highlights the

https://doi.org/10.1192/bjp.180.1.13 Published online by Cambridge University Press

importance of design and execution issues if a CBA study is to produce interpretable findings. The fundamental problems with the design of this study – a poorly matched control group and non-contemporaneous data collection in intervention and control groups (Marshall *et al*, 1999) – led its critics to call for a well-designed RCT in a UK setting. Although a (cluster) RCT may well be desirable, there is no reason why a well-designed CBA could not provide some of the answers on the effectiveness of community mental health teams and case management in a UK setting.

The CBA design has been most convincingly used in US mental health services research, where large administrative and quality assurance databases exist of patient outcomes collected routinely on a service-wide basis. These have been used, for example, to evaluate case management (Lam & Rosenheck, 1999) and the adverse consequences of mental health reimbursement cutbacks in US insurance plans (Rosenheck *et al*, 1999). However, in the UK, characterising and outcome data are rarely collected routinely. If similarly sophisticated information systems do become available in the UK, then there is the potential for cohort and CBA studies to become easier to conduct.

Interrupted time series studies

In the ITS design, multiple observations are made within a single group before and after an intervention is implemented (Cook & Campbell, 1979). In this case, the 'control' observations are those made before the intervention is implemented (Ukoumunne *et al*, 1999*b*). Because the strength of the design is in the ability to take trends over time into account, as a general rule, studies should include more than three observations both before and after an intervention to give interpretable results (Ukoumunne *et al*, 1999*b*).

Time series analyses are particularly useful for evaluating the effects of an intervention when it is not possible to identify an appropriate control group (Campbell *et al*, 1999). This study design has been used to evaluate the impact of wide dissemination of information on the more rational use of newer antidepressants (Mason *et al*, 1998) and also could be used to evaluate the impact on prescribing and practice of the dissemination of national guidelines, such as those planned by the National Institute for Clinical Excellence (NICE) on atypical antipsychotic medication. Time series analyses can detect whether an intervention has had an effect significantly greater than the underlying trend (Cook & Campbell, 1979). The main problem with the design is that it does not protect against the effects of other events occurring at the same time as the intervention of interest, which could also affect performance on the study measures. However, there are several variants on the basic design that can help to overcome this problem, including (Cook & Campbell, 1979):

(a) where it is possible to include a control group in the time series;

(b) measuring other variables over time to indicate the likely influence, or not, of other events;

(c) 'staggered' time series (where this is possible).

The statistical analysis of ITSs is particularly complex, however, and interested readers are advised to seek expert help if they want to carry out this type of study.

Conclusions

In summary, the strength of quasi-experimental study designs lies in their 'real-world' applicability: not only may they be the only possible evaluation design in many circumstances, but their findings may be more generalisable. The key is to use study designs that produce interpretable findings, such as those described here, and to avoid, for example, studies of single groups without control groups or without repeated pre- and post-intervention observations (Ukoumunne *et al*, 1999*b*). However, even in the best designed and executed studies, where potential confounders are controlled for in the design and analysis, the possibility of bias from residual or unknown confounders remains. For this reason, the gold-standard evaluation design for organisational interventions is the clustered RCT and a quasi-experimental design should be undertaken only when its use can be justified.

DISCUSSION

There are important questions in mental health regarding the best methods of implementing services or improving clinical care. The NSF for mental health (Secretary of State for Health, 1999) makes explicit that the implementation of this framework will be shaped by "evidence of clinical and cost effectiveness of mental health services". Such evidence will come only from appropriately designed evaluations of policy or implementation strategies conducted in routine care settings. Cluster-based studies remain the gold standard in evaluating many organisational and policy initiatives (Ukoumunne *et al*, 1999*b*) and should form the cornerstone of the evidence base for mental health care. However, cluster-based studies may be neither feasible nor possible, particularly where implementation precedes evaluation, as is the case in the NSF (Tyrer, 1999) and with other policy initiatives in other health service areas (Aday *et al*, 1998).

It is clear that mental health policy will continue to be made in the absence of a clear evidence base, and a pragmatic response of researchers should be an ability to use experimental designs that may be neither their first choice nor the most robust. There are settings where quasi-experimental study designs may be the only option for evaluating many of the organisational issues in the NSF. It is timely that specific funds have been allocated alongside the NSF to set up an SDO research and development programme, with a specific remit to examine mental health (Secretary of State for Health, 1999). It remains to be seen whether governments and health policy makers will be prepared to act on the basis of this research, in order to ensure that health policy is truly shaped by evidence. Clearly this evidence will not be in the form of the conventional individualised RCT, where this has hitherto been regarded as the gold standard in the hierarchy of evidence. The problem remains that health care decisions often are not based on the best evidence (randomised or otherwise), and translating evidence into practice and policy remains a challenge for all health care (NHS Centre for Reviews and Dissemination, 1999; Lawrie *et al*, 2001).

APPENDIX

Intraclass correlation and the design effect

A technical definition of the intraclass correlation (ρ) is the proportion of the true variation in the outcome that can be attributed to differences between clusters:

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

where σ_b^2 is the between-cluster variance component and σ_w^2 is the within-cluster variance component.

In order to take account of between-cluster variation in sample size estimation, hypothesis testing and confidence interval estimation, the variance term

https://doi.org/10.1192/bjp.180.1.13 Published online by Cambridge University Press

in standard statistical formulae must be increased by the design effect. The most commonly used formula for design effect estimation is given by:

$$\text{Design effect} = 1 + (n - 1)\rho$$

where n is the average cluster size. It follows from this equation that the larger the average cluster size, the larger the design effect. In the case of sample size estimation, conventional sample size calculations will need to be inflated by the design effect. Similarly, for large clusters, such as health authority population or hospital catchment areas, the design effect will be large even for very low levels of intraclass correlation. For a useful technical summary, see Ukoumunne *et al* (1999*b*).

CLINICAL IMPLICATIONS

- This article highlights the need to provide an evidence base for mental health policy, service delivery and organisation.
- Cluster randomised trials form the gold standard of this research evidence.
- The importance of clustering often is not appreciated in the design and analysis of mental health services research.

LIMITATIONS

- Policy implementation often precedes evaluation, making cluster randomised studies difficult to conduct.
- Quasi-experimental studies are an alternative, but their limitations need to be taken into account.
- The ability of quasi-experimental studies to provide robust evidence in mental health services research remains to be examined.

SIMON GILBODY, MRCPsych, Academic Unit of Psychiatry and Behavioural Sciences, University of Leeds, Leeds; PAULA WHITTY, MFPHM, Department of Epidemiology and Public Health, University of Newcastle upon Tyne, Newcastle upon Tyne, UK

Correspondence: Simon Gilbody, Academic Unit of Psychiatry and Behavioural Sciences, University of Leeds, Leeds LS2 9LT, UK. Tel: 0113 233 1899; fax: 0113 243 3719; e-mail: s.m.gilbody@leeds.ac.uk

(First received 6 November 2000, final revision 12 April 2001, accepted 20 April 2001)

REFERENCES

Aday, L. A., Begley, C. E., Lairson, D. R., *et al* (1998) *Evaluating Healthcare System: Effectiveness, Efficiency and Equity* (2nd edn). Chicago, IL: AHSR.

Begg, C., Cho, M., Eastwood, S., *et al* (1996) Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *Journal of the American Medical Association*, **276**, 637–639.

Black, N. (1996) Why we need observational studies to evaluate the effectiveness of health care. *British Medical Journal*, **312**, 1215–1218.

___, Brazier, J., Fitzpatrick, R., *et al* (1998) *Health Services Research Methods: a Guide to Best Practice*. London: BMJ Press.

Buck, C. & Donner, A. (1982) The design of controlled experiments in the evaluation of non-therapeutic interventions. *Journal of Chronic Diseases*, **35**, 531–538.

Campbell, D. T. & Stanley, J. C. (1966) *Experimental and Quasi-experimental Designs for Research*. Chicago: Rand McNally.

Campbell, M. K. & Grimshaw, J. M. (1998) Cluster randomised trials: time for improvement. *British Medical Journal*, **317**, 1171.

___, Steen, N., Grimshaw, J. M., *et al* (1999) Design and statistical issues in implementation research. In *Changing Professional Practice. Theory and Practice of Clinical Guidelines Implementation* (ed. M. M. Thorsen), pp. 57–76. Copenhagen: Danish Institute for Health Services Research and Development.

___, Fitzpatrick, R., Haines, A., *et al* (2000*a*) Framework for design and evaluation of complex interventions to improve health. *British Medical Journal*, **321**, 694–696.

___, Mollinson, J., Steen, N., *et al* (2000*b*) Analysis of cluster randomised trials in primary care: a practical approach. *Family Practice*, **17**, 192–196.

Cook, T. D. & Campbell, D. T. (1979) *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Boston, MA: Houghton Mifflin.

Department of Health (2000) *Research and Development for a First Class Service: R&D Funding in the New NHS*. London: Department of Health.

Divine, G. W., Brown, J. T. & Frazer, L. M. (1992) The unit of analysis error in studies about physicians' patient care behavior. *Journal of General Internal Medicine*, **7**, 623–629.

Donner, A. & Klar, N. (1994) Methods for comparing event rates in intervention studies when the unit of allocation is a cluster. *American Journal of Epidemiology*, **140**, 279–289, 300–301.

___, Brown, K. S. & Brasher, P. (1990) A methodological review of non-therapeutic intervention trials employing cluster randomization, 1979–1989. *International Journal of Epidemiology*, **19**, 795–800.

Elbourne, D. (1997) Guidelines are needed for evaluations that use a cluster approach. *British Medical Journal*, **315**, 1620–1621.

Fairhurst, K. & Dorwick, C. (1996) Problems with recruitment in a randomised trial of counselling in general practice: causes and implications. *Journal of Health Services Research and Policy*, **1**, 77–80.

Family Heart Study Group (1994) Randomised controlled trial evaluating cardiovascular screening and intervention in general practice: principal results of British family heart study. *British Medical Journal*, **308**, 313–320.

Gail, M. H., Mark, S. D., Carroll, R. J., *et al* (1996) On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine*, **15**, 1069–1092.

Gilbody, S. M., House, A. O. & Sheldon, T. A. (2001) Routinely administered questionnaires for depression and anxiety: a systematic review. *British Medical Journal*, **322**, 406–409.

Goldberg, D. P. & Williams, P. (1988) *The User's Guide to the General Health Questionnaire*. Windsor: NFER-Nelson.

Hawton, K. & James, R. (1995) General hospital services for attempted suicide: a survey of one region. *Health Trends*, **27**, 18–21.

___, Arensman, E., Townsend, E., *et al* (1998) Deliberate self harm: systematic review of efficacy of psychosocial and pharmacological treatments in preventing repetition. *British Medical Journal*, **317**, 441–447.

Hoeper, E. W., Nycz, G. R., Kessler, L. G., *et al* (1984) The usefulness of screening for mental illness. *Lancet*, *i*, 33–35.

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.

Jacobs, D., Luepker, R. V. & Mittelmark, M. (1986) Community-wide prevention strategies: evaluation design of the Minnesota Heart Health program. *Journal of Chronic Diseases*, **39**, 775–788.

Johnstone, A. & Goldberg, D. (1976) Psychiatric screening in general practice. A controlled trial. *Lancet*, *i*, 605–608.

Katon, W., Von Korff, M., Lin, E., *et al* (1995) Collaborative management to achieve treatment guidelines. Impact on depression in primary care. *Journal of the American Medical Association*, **273**, 1026–1031.

Kerry, S. M. & Bland, J. M. (1998*a*) The intracluster correlation coefficient in cluster randomisation. *British Medical Journal*, **316**, 1455.

___ & ___ (1998*b*) Trials which randomize practices. I: How should they be analysed? *Family Practice*, **15**, 80–83.

https://doi.org/10.1192/bjp.180.1.13 Published online by Cambridge University Press

__ & __ (1998*c*) Trials which randomize practices. II: Sample size. *Family Practice*, **15**, 84–87.

Lam, J. A. & Rosenheck, R. (1999) Street outreach for homeless persons with serious mental illness: is it effective? *Medical Care*, **37**, 894–907.

Lawrie, S. M., Scott, A. I. F. & Sharpe, M. C. (2001) Implementing evidence-based psychiatry: whose responsibility? *British Journal of Psychiatry*, **178**, 195–196.

Marshall, M., Bond, G., Stein, L. I., *et al* (1999) PRiSM Psychosis Study. Design limitations, questionable conclusions. *British Journal of Psychiatry*, **175**, 501–503.

Mason, J., Freemantle, N. & Young, P. (1998) The effect of the distribution of Effective Health Care Bulletins on prescribing selective serotonin reuptake inhibitors in primary care. *Health Trends*, **30**, 120–123.

McKee, M., Britton, A., Black, N., *et al* (1998) Choosing between randomised and non-randomised studies. In *Health Services Research Methods: a Guide to Best Practice* (eds N. Black, J. Brazier, R. Fitzpatrick, *et al*), pp. 61–72. London: BMJ Press.

Meinert, C. L. (1986) *Clinical Trials: Design, Conduct and Analysis*. Oxford: Oxford University Press.

National Co-ordinating Centre for NHS Service Delivery and Organisation Research and Development (2000) *Using Research to Improve Health Care Services*. London: London School of Hygiene and Tropical Medicine.

NHS Centre for Reviews and Dissemination (1994) *Implementing Practice Guidelines*. (Vol. 1). York: University of York.

__ (1999) *Getting Evidence into Practice* (Vol. 4). York: University of York.

Rand, E., Badger, L. & Cogins, D. (1988) Towards a resolution of contradictions: utility of feedback from the GHQ. *General Hospital Psychiatry*, **10**, 189–196.

Reeves, B., Maclehose, R., Harvey, I. M., *et al* (1998) Comparison of effect sizes derived from randomised and non randomised studies. In *Health Services Research Methods: a Guide to Best Practice* (eds N. Black, J. Brazier, R. Fitzpatrick, *et al*), pp. 73–85. London: BMJ Press.

Rice, N. & Leyland, A. (1996) Multi-level models: applications to health data. *Journal of Health Services Research and Policy*, **3**, 154–164.

Rosenheck, R. A., Druss, B., Stolar, M., *et al* (1999) Effect of declining mental health service use on employees of a large corporation. *Health Affairs*, **18**, 193–203.

Sackett, D. L., Haynes, R. B., Guyatt, G. H., *et al* (1991) *Clinical Epidemiology: a Basic Science for Clinical Medicine*. Boston, MA: Little, Brown and Company.

Secretary of State for Health (1999) *National Service Framework for Mental Health: Modern Standards and Service Models*. London: Department of Health.

Simpson, J. M., Klar, N. & Donnor, A. (1995) Accounting for cluster randomization: a review of primary prevention trials, 1990 through 1993. *American Journal of Public Health*, **85**, 1378–1383.

Smith, P., Sheldon, T. A. & Martin, S. (1996) An index of need for psychiatric services based on in-patient utilisation. *British Journal of Psychiatry*, **169**, 308–321.

Thompson, C., Kinmonth, J., Stevens, L., *et al* (2000) Effects of a clinical-practice guideline and practice-based education on detection and outcome of depression in primary care: Hampshire Depression Project randomised controlled trial. *Lancet*, **355**, 50–57.

Thornicroft, G., Strathdee, G., Phelan, M., *et al* (1998) Rationale and design. PRiSM Psychosis Study I. *British Journal of Psychiatry*, **173**, 363–370.

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

Tyrer, P. (1999) The national service framework: a scaffold for mental health. *British Medical Journal*, **319**, 1017–1018.

Ukoumunne, O. C., Gulliford, M. C., Chinn, S., *et al* (1999*a*) Methods in health service research. Evaluation of health interventions at area and organisation level. *British Medical Journal*, **319**, 376–379.

__, __, __, *et al* (1999*b*) Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technology Assessment*, **3**, iii–92.

Wells, K. B. (1999) Treatment research at the crossroads: the scientific interface of clinical trials and effectiveness research. *American Journal of Psychiatry*, **156**, 5–10.

Whiting-O'Keefe, Q. E., Henke, C. & Simborg, D. W. (1984) Choosing the correct unit of analysis in medical care experiments. *Medical Care*, **22**, 1101–1114.

World Health Organization (1991) *Evaluation of Methods for the Treatment of Mental Disorders*. Geneva: WHO.

https://doi.org/10.1192/bjp.180.1.13 Published online by Cambridge University Press