QUALITY IMPROVEMENT RESEARCH

Research designs for studies evaluating the effectiveness of change and improvement strategies

M Eccles, J Grimshaw, M Campbell, C Ramsay

Qual Saf Health Care 2003;**12**:47–52

The methods of evaluating change and improvement strategies are not well described. The design and conduct of a range of experimental and non-experimental quantitative designs are considered. Such study designs should usually be used in a context where they build on appropriate theoretical, qualitative and modelling work, particularly in the development of appropriate interventions. A range of experimental designs are discussed including single and multiple arm randomised controlled trials and the use of more complex factorial and block designs. The impact of randomisation at both group and individual levels and three non-experimental designs (uncontrolled before and after, controlled before and after, and time series analysis) are also considered. The design chosen will reflect both the needs (and resources) in any particular circumstances and also the purpose of the evaluation. The general principle underlying the choice of evaluative design is, however, simple—those conducting such evaluations should use the most robust design possible to minimise bias and maximise generalisability.

See end of article for authors' affiliations

Correspondence to: Professor M Eccles, Professor of Clinical Effectiveness, Centre for Health Services Research, 21 Claremont Place, Newcastle upon Tyne NE2 4AA, UK; martin.eccles@ncl.ac.uk

Accepted 29 October 2002

There is a substantial literature about the design, conduct, and analysis of evaluations of relatively simple healthcare interventions such as drugs. However, the methods of evaluating complex interventions such as quality improvement interventions are less well described. Evaluation informs the choice between alternative interventions or policies by identifying, estimating and, if possible, valuing the advantages and disadvantages of each.[1]

There are a number of quantitative designs that could be used to evaluate quality improvement interventions (box 1).

All of these designs attempt to establish general causal relationships across a population of interest. The choice of design will be dependent upon the purpose of the evaluation and the degree of control the researchers have over the delivery of the intervention(s). In general, researchers should choose a design that minimises potential bias (any process at any stage of inference which tends to produce results or conclusions that differ systematically from the truth; also referred to as internal validity) and maximises generalisability (the degree to which the results of a study hold true for situations other than those pertaining to the study, in particular for routine clinical practice; also referred to as external validity).[2 3]

A FRAMEWORK FOR EVALUATING QUALITY IMPROVEMENT INTERVENTIONS

Campbell and colleagues[4] have suggested that the evaluation of complex interventions should follow a sequential approach involving:

- development of the theoretical basis for an intervention;
- definition of components of the intervention (using modelling, simulation techniques or qualitative methods);
- exploratory studies to develop further the intervention and plan a definitive evaluative study (using a variety of methods);
- definitive evaluative study (using quantitative evaluative methods, predominantly randomised designs).

This framework demonstrates the interrelation between quantitative evaluative methods and other methods; it also makes explicit that the design and conduct of quantitative evaluative studies should build upon the findings of other quality improvement research. However, it represents an idealised framework and, in some circumstances, it is necessary to undertake evaluations without sequentially working through the earlier stages—for example, when evaluating policy interventions that are being introduced without prior supporting evidence.

In this paper we describe quantitative approaches for evaluating quality improvement interventions, focusing on methods for estimating the magnitude of the benefits. We also focus on the evaluation of interventions within systems rather than evaluations of whole systems. We discuss several study designs for definitive evaluative studies including a range of randomised controlled trial designs and three non-randomised or quasi-experimental evaluative designs.

Box 1 Possible quantitative evaluative designs for quality improvement research

Randomised designs

- Individual patient randomised controlled trials
- Cluster randomised trials

Non-randomised designs

- Uncontrolled before and after studies
- Controlled before and after studies
- Time series designs

EVALUATIVE DESIGNS

Randomised designs

Randomised trials are the gold standard method for evaluating healthcare interventions.[5] They estimate the impact of an intervention through direct comparison with a randomly allocated control group that either receives no intervention or an alternative intervention.[6] The randomisation process is the best way of ensuring that both known and (particularly importantly) unknown factors (confounders) that may independently affect the outcome of an intervention are likely to be distributed evenly between the trial groups. As a result, differences observed between groups can be more confidently ascribed to the effects of the intervention rather than to other factors. The same arguments that are used to justify randomised controlled trials of clinical interventions such as drugs are at least as salient to the evaluations of quality improvement interventions. In particular, given our incomplete understanding of potential confounders relating to organisational or professional performance, it is even more difficult to adjust for these in non-randomised designs.

Cluster randomisation

While it is possible to conduct randomised trials of quality improvement interventions which randomise individual patients, this may not always be ideal. If there is the possibility that the treatment given to control individuals will be affected by an organisation's or professional's experience of applying the intervention to other patients in the experimental group, there is a risk of contamination. For example, Morgan *et al*[7] investigated the effects of computerised reminders for antenatal care. Patients were randomised and physicians received reminders for intervention patients but not control patients. Compliance in intervention patients rose from 83% to 98% over 6 months, while compliance in control patients rose from 83% to 94% over 12 months. This is a probable contamination effect.

If such contamination is likely, the researcher should consider randomising organisations or healthcare professionals rather than individual patients, although data may still be collected about the process and outcome of care at the individual patient level. Such trials, which randomise at one level (organisation or professional) and collect data at a different level (patient), are known as cluster randomised trials.[8 9] Cluster randomisation has considerable implications for the design, power, and analysis of studies which have frequently been ignored.

Design considerations

The main design considerations concern the level of randomisation and whether to include baseline measurement. Frequently researchers need to trade off the likelihood of contamination at lower levels of randomisation against decreasing numbers of clusters and increasing logistical problems at higher levels of randomisation. For example, in a study of an educational intervention in secondary care settings, potential levels of randomisation would include the individual clinician, the ward, the clinical service or directorate, and the hospital. Randomisation at the level of the hospital would minimise the risk of contamination but dramatically increase the size and complexity of the study due to the greater number of hospitals required. Randomisation at the level of the individual clinician would decrease the number of hospitals required but there may then be a risk of contamination across clinicians working in the same wards or specialty areas.

In situations where relatively few clusters (e.g. hospitals) are available for randomisation, there is increased danger of imbalance in performance between study and control groups due to the play of chance. Baseline measurements can be used to assess adequacy of the allocation process and are also useful because they provide an estimate of the magnitude of a problem. Low performance scores before the intervention may indicate that performance is poor and there is much room for improvement, whereas high performance scores may indicate that there is little room for improvement (ceiling effect). In addition, baseline measures could be used as a stratifying or matching variable or incorporated in the analysis to increase statistical power (see below). These potential benefits have to be weighed against the increased costs and duration of studies incorporating baseline measurements and concerns about testing effects (introduction of potential bias due to sensitisation of the study subjects during baseline measurement).[2]

Sample size calculation

A fundamental assumption of the standard statistics used to analyse patient randomised trials is that the outcome for an individual patient is completely unrelated to that for any other patient—they are said to be "independent". This assumption is violated, however, when cluster randomisation is adopted because two patients within any one cluster are more likely to respond in a similar manner than are two patients from different clusters. For example, the management of patients in a single hospital is more likely to be consistent than management of patients across a number of hospitals. The primary consequence of adopting a cluster randomised design is that it is not as statistically efficient and has lower statistical power than a patient randomised trial of equivalent size.

Sample sizes for cluster randomised trials therefore need to be inflated to adjust for clustering. A statistical measure of the extent of clustering is known as the "intracluster correlation coefficient" (ICC) which is based on the relationship of the between-cluster to within-cluster variance.[10] Table 1 shows a number of ICCs from a primary care study of computerising guidelines for patients with either asthma or stable angina (box 4).

Both the ICC and the cluster size influence the inflation required; the sample size inflation can be considerable especially if the average cluster size is large. The extra numbers of patients required can be achieved by increasing either the number of clusters in the study (the more efficient method[11]) or the number of patients per cluster. In general, little additional power is gained from increasing the number of patients per cluster above 50. Researchers often have to trade off the logistical difficulties and costs associated with recruitment of extra clusters against those associated with increasing the number of patients per cluster.[12]

Analysis of cluster randomised trials

There are three general approaches to the analysis of cluster randomised trials: analysis at cluster level; the adjustment of standard tests; and advanced statistical techniques using data recorded at both the individual and cluster level.[9 13 14] Cluster level analyses use the cluster as the unit of randomisation and analysis. A summary statistic (e.g. mean, proportion) is computed for each cluster and, as each cluster provides only one data point, the data can be considered to be independent, allowing standard statistical tests to be used. Patient level analyses can be undertaken using adjustments to simple statistical tests to account for the clustering effect. However, this approach does not allow adjustment for patient or practice characteristics. Recent advances in the development and use of new modelling techniques to incorporate patient level data allow the inherent correlation within clusters to be modelled explicitly, and thus a "correct" model can be obtained. These methods can incorporate the hierarchical nature of the data into the analysis. For example, in a primary care setting we may have patients (level 1) treated by general practitioner (level 2) nested within practices (level 3) and may have covariates measured at the patient level (e.g. patient age or sex), the general practitioner level (e.g. sex, time in practice), and at the practice level (e.g. practice size). Which of the

Table 1 ICCs for medical record and prescribing data.

Angina		Asthma	
Process of care measures			
Number of consultations	0.04	Number of consultations	0.03
Number of consultations for angina	0.16	Number of consultations for asthma	0.05
Was blood pressure recorded?	0.04	Compliance checked?	0.15
Was exercise level or advice about exercise recorded?	0.08	Inhaler technique checked?	0.10
Any advice about Mediterranean diet or oily fish?	0.01	Lung function recorded?	0.08
Weight or advice about weight recorded?	0.10	Asthma education recorded?	0.12
Smoking status recorded?	0.10	Smoking status recorded?	0.09
		Smoking advice/education recorded?	0.03
ECG recorded?	0.01		
Thyroid function recorded?	0.01		
Blood glucose or HbA1c recorded?	0.03		
Cholesterol or other lipids recorded?	0.04		
Haemoglobin recorded?	0.05		
Exercise ECG recorded?	0.02		
Drugs			
Was verapamil prescribed?	0.01	Was a short acting β_2 agonist prescribed?	0.02
Was a beta blocker prescribed?	0.01	Inhaled corticosteroids	0.02
Short acting GTN	0.01	Long acting β_2 agonists	0.01
Modified release GTN	0.01	Oral steroids	0.02
Transdermal GTN	0.02	Oral bronchodilators	0.02
Isosorbide dinitrate (SA & MR)	0.07	Prescribing of inhaled corticosteroids for subjects who were	0.04
Isosorbide mononitrate (SA & MR)	0.02	prescribed a mean daily dose of >6 puffs	
Diltiazem	0.02		
Ca channel blocker	0.01		
Statins	0.02		
Beta blocker and dinitrate	0.04		
Calcium blocker and dinitrate	0.04		
Nitrate, calcium blocker and β blocker	0.02		

methods is better to use is still a topic of debate. The main advantage of such sophisticated statistical methods is their flexibility. However, they require extensive computing time and statistical expertise, both for the execution of the procedures and in the interpretation of the results.

No consensus exists as to which approach should be used. The most appropriate analysis option will depend on a number of factors including the research question; the unit of inference; the study design; whether the researchers wish to adjust for other relevant variables at the individual or cluster level (covariates); the type and distribution of outcome measure; the number of clusters randomised; the size of cluster and variability of cluster size; and statistical resources available in the research team. Campbell *et al*[15] and Mollison *et al*[16] present worked examples comparing these different analytical strategies.

Possible types of cluster randomised trials

Two arm trials

The simplest randomised design is the two arm trial where each subject is randomised to study or control groups. Observed differences in performance between the groups are assumed to be due to the intervention. Such trials are relatively straightforward to design and conduct and they maximise statistical power (half the sample is allocated to the intervention and half to the control). However, they only provide information about the effectiveness of a single intervention compared with control (or the relative effectiveness of two interventions without reference to a control). Box 2 shows an example of a two arm trial.

Box 2 Two arm trial[17]

The trial aimed to assess whether the quality of cardiovascular preventive care in general practice could be improved through a comprehensive intervention implemented by an educated outreach visitor. After baseline measurements, 124 general practices (in the southern half of the Netherlands) were randomly allocated to either intervention or control. The intervention, based on the educational outreach model, comprised 15 practice visits over a period of 21 months and addressed a large number of issues around task delegation, availability of instruments and patient leaflets, record keeping, and follow up routines. Twenty one months after the start of the intervention, post-intervention measurements were performed. The difference between ideal and actual practice in each aspect of organising preventive care was defined as a deficiency score. The primary outcome measure was the difference in deficiency scores before and after the intervention. All practices completed both baseline and post-intervention measurements. The difference in change between intervention and control groups, adjusted for baseline, was statistically significant ($p<0.001$) for each aspect of organising preventive care. The largest absolute improvement was found for the number of preventive tasks performed by the practice assistant.

Multiple arm trials

The simplest extension to the two arm trial is to randomise groups of professionals to more than two groups—for example, two or more study groups and a control group. Such studies are relatively simple to design and use, and allow head-to-head comparisons of interventions or levels of intervention under similar circumstances. These benefits are, however, compromised by a loss of statistical power; for example, to achieve the same power as a two arm trial, the sample size for a three arm trial needs to be increased by up to 50%.

Factorial designs

Factorial designs allow the evaluation of the relative effectiveness of more than one intervention compared with control. For example, in a 2 × 2 factorial design evaluating two interventions against control, participants are randomised to each intervention (A and B) independently. In the first

Box 3 Factorial design[19]

The trial evaluated the effectiveness of audit and feedback and educational reminder messages to change general practitioners' radiology ordering behaviour for lumber spine and knee *x* rays. The design was a before and after pragmatic cluster randomised controlled trial using a 2×2 factorial design involving 244 practices and six radiology departments in two geographical regions. Each practice was randomised twice, to receive or not each of the two interventions. Educational reminder messages were based on national guidelines and were provided on the report of every relevant *x* ray ordered during the 12 month intervention period. For example, the lumbar spine message read "In either acute (less than 6 weeks) or chronic back pain, without adverse features, *x* ray is not routinely indicated". The audit and feedback covered the preceding 6 month period and was delivered to individual general practitioners at the start of the intervention period and again 6 months later. It provided practice level information relating the number of requests made by the whole practice relative to the number of requests made by all practices in the study. Audit and feedback led to a non-significant reduction of around 1% *x* ray requests while educational reminder messages led to a relative reduction of about 20% *x* ray requests.

Box 4 Balanced incomplete block design[20]

This study was a before and after pragmatic cluster randomised controlled trial using a 2×2 incomplete block design and was designed to evaluate the use of a computerised decision support system (CDSS) in implementing evidence based clinical guidelines for the primary care management of asthma in adults and angina. It was based in 60 general practices in the north east of England and the participants were general practitioners and practice nurses in the study practices and their patients aged 18 years or over and with angina or asthma. The practices were randomly allocated to two groups. The first group received computerised guidelines for the management of angina and provided intervention patients for the management of angina and control patients for the management of asthma. The second received computerised guidelines for the management of asthma and provided intervention patients for the management of asthma and control patients for the management of angina. The outcome measures were adherence to the guidelines, determined by recording of care in routine clinical records, and any subsequent impact measured by patient reported generic and condition specific measures of outcome. There were no significant effects of CDSS on consultation rates, process of care measures (including prescribing), or any quality of life domain for either condition. Levels of use of the CDSS were low.

randomisation the study participants are randomised to intervention A or control. In the second randomisation the same participants are randomised to intervention B or control. This results in four groups: no intervention, intervention A alone, intervention B alone, interventions A and B.

During the analysis of factorial designs it is possible to undertake independent analyses to estimate the effect of the interventions separately[18]; essentially this design allows the conduct of two randomised trials for the same sample size as a two arm trial. However, these trials are more difficult to operationalise and analyse, they provide only limited power for a direct head-to-head comparison of the two interventions, and the power is diminished if there is interaction between the two interventions. Box 3 shows an example of a factorial design trial that was powered to be able to detect any interaction effects.

Balanced incomplete block designs

In guideline implementation research there are a number of non-specific effects which may influence the estimate of the effect of an intervention. These could be positive attention effects from participants knowing that they are the subject of a study, or negative demotivation effects from being allocated to a control rather than an intervention group. Currently, these non-specific effects are grouped together and termed the "Hawthorne effect". If these are imbalanced across study groups in a quality improvement trial, the resulting estimates of effects may be biased and, as these effects can potentially be of the same order of magnitude as the effects that studies are seeking to demonstrate, there is an advantage to dealing with them systematically. While these effects may difficult to eliminate, balanced incomplete block designs can be used to equalise such non-specific effects and thereby minimise their impact.[18] An example is shown in box 4.

As doctors in both groups were subject to the same level of intervention, any non-specific effects are equalised across the two groups leaving any resulting difference as being due to the intervention.

Non-randomised designs

Quasi-experimental designs

Quasi-experimental designs are useful where there are political, practical, or ethical barriers to conducting a genuine (randomised) experiment. Under such circumstances, researchers have little control over the delivery of an intervention and have to plan an evaluation around a proposed intervention. A large number of potential designs have been summarised by Campbell and Stanley[2] and Cook and Campbell.[3] Here we discuss the three most commonly used designs in quality improvement studies: (1) uncontrolled before and after studies, (2) controlled before and after studies, and (3) time series designs.

Uncontrolled before and after studies

Uncontrolled before and after studies measure performance before and after the introduction of an intervention in the same study site(s) and observed differences in performance are assumed to be due to the intervention. They are relatively simple to conduct and are superior to observational studies, but they are intrinsically weak evaluative designs because secular trends or sudden changes make it difficult to attribute observed changes to the intervention. There is some evidence to suggest that the results of uncontrolled before and after studies may overestimate the effects of quality improvement-like interventions. Lipsey and Wilson[21] undertook an overview of meta-analyses of psychological, educational and behavioural interventions. They identified 45 reviews that reported separately the pooled estimates from controlled and uncontrolled studies, and noted that the observed effects from uncontrolled studies were greater than those from controlled studies. In general, uncontrolled before and after studies should not be used to evaluate the effects of quality improvement interventions and the results of studies using such designs have to be interpreted with great caution.

Controlled before and after studies

In controlled before and after studies the researcher attempts to identify a control population of similar characteristics and

Box 5 Time series analysis[22]

An interrupted time series using monthly data for 34 months before and 14 months after dissemination of the guidelines was used to evaluate the effect of postal dissemination of the third edition of the Royal College of Radiologists' guidelines on general practitioner referrals for radiography. Data were abstracted for the period April 1994 to March 1998 from the computerised administrative systems of open access radiological services provided by two teaching hospitals in one region of Scotland. A total of 117 747 imaging requests from general practice were made during the study period. There were no significant effects of disseminating the guidelines on the total number of requests or 18 individual tests. If a simple before and after study was used, then we would have erroneously concluded that 11 of the 18 procedures had significant differences.

performance to the study population and collects data in both populations before and after the intervention is applied to the study population. Analysis compares post-intervention performance or change scores in the study and control groups and observed differences are assumed to be due to the intervention.

While well designed before and after studies should protect against secular trends and sudden changes, it is often difficult to identify a comparable control group. Even in apparently well matched control and study groups, performance at baseline often differs. Under these circumstances, "within group" analyses are often undertaken (where change from baseline is compared within both groups separately and where the assumption is made that, if the change in the intervention group is significant and the change in the control group is not, the intervention has had an effect). Such analyses are inappropriate for a number of reasons. Firstly, the baseline imbalance suggests that the control group is not truly comparable and may not experience the same secular trends or sudden changes as the intervention group; thus any apparent effect of the intervention may be spurious. Secondly, there is no direct comparison between study and control groups.[2] Another common analytical problem in practice is that researchers fail to recognise clustering of data when interventions are delivered at an organisational level and data are collected at the individual patient level.

Time series designs

Time series designs attempt to detect whether an intervention has had an effect significantly greater than the underlying secular trend.[3] They are useful in quality improvement research for evaluating the effects of interventions when it is difficult to randomise or identify an appropriate control group—for example, following the dissemination of national guidelines or mass media campaigns (box 5). Data are collected at multiple time points before and after the intervention. The multiple time points before the intervention allow the underlying trend and any cyclical (seasonal) effects to be estimated, and the multiple time points after the intervention allow the intervention effect to be estimated while taking account of the underlying secular trends.

The most important influence on the analysis technique is the number of data points collected before the intervention. It is necessary to collect enough data points to be convinced that a stable estimate of the underlying secular trend has been obtained. There are a number of statistical techniques that can be used depending on the characteristics of the data, the number of data points available, and whether autocorrelation is present,[3] Autocorrelation refers to the situation whereby data points collected close in time are likely to be more similar to each other than to data points collected far apart. For example, for any given month the waiting times in hospitals are likely to be more similar to waiting times in adjacent months than to waiting times 12 months previously. Autocorrelation has to be allowed for in analysis and time series regression models,[23] and autoregressive integrated moving averages (ARIMA) modelling[3] and time series regression models[23] are all methods for dealing with this problem.

Well designed time series evaluations increase the confidence with which the estimate of effect can be attributed to the intervention, although the design does not provide protection against the effects of other events occurring at the same time as the study intervention, which might also improve performance. Furthermore, it is often difficult to collect sufficient data points unless routine data sources are available. It has been found that many published time series studies have been inappropriately analysed, frequently resulting in an overestimation of the effect of the intervention.[24 25]

DISCUSSION

Randomised trials should only be considered when there is genuine uncertainty about the effectiveness of an intervention. Whilst they are the optimal design for evaluating quality improvement interventions, they are not without their problems. They can be logistically difficult, especially if the researchers are using complex designs to evaluate more than one intervention or if cluster randomisation—requiring the recruitment of large numbers of clusters—is planned. They are undoubtedly methodologically challenging and require a multidisciplinary approach to adequately plan and conduct. They can also be time consuming and expensive; in our experience a randomised trial of a quality improvement intervention can rarely be completed in less than 2 years.

Critics of randomised trials frequently express concerns that tight inclusion criteria of trials or artificial constraints placed upon participants limit the generalisability of the findings. While this is a particular concern in efficacy (explanatory) studies of drugs, it is likely to be less of a problem in quality improvement evaluations that are likely to be inherently pragmatic.[26] Pragmatic studies aim to test whether an intervention is likely to be effective in routine practice by comparing the new procedure against the current regimen; as such they are the most useful trial design for developing policy recommendations. Such studies attempt to approximate normal conditions and do not attempt to equalise contextual factors and other effect modifiers in the intervention and study groups. In pragmatic studies, the contextual and effect modifying factors therefore become part of the interventions. Such studies are usually conducted on a predefined study population and withdrawals are included within an "intention to treat" analysis; all subjects initially allocated to the intervention group would be analysed as intervention subjects irrespective of whether they received the intervention or not. For example, in an evaluation of a computerised decision support system as a method of delivering clinical guidelines in general practice (box 4), some general practitioners may not have had the computing skills to work the intervention. In an intention to treat analysis, data from all general practitioners would be included in the analysis irrespective of whether they could use the system or not; as a result, the estimates of effect would more likely reflect the effectiveness of the intervention in real world settings.

The main limitation of quasi-experimental designs is that the lack of randomised controls threatens internal validity and increases the likelihood of plausible rival hypotheses. Cook and Campbell[3] provide a framework for considering the internal validity of the results of experiments and quasi experiments when trying to establish causality. They suggest that "Estimating the internal validity of a relationship is a

Key messages

- Whatever design is chosen, it is important to minimise bias and maximise generalisability.
- Quantitative designs should be used within a sequence of evaluation building as appropriate on preceding theoretical, qualitative, and modelling work.
- There are a range of more or less complex randomised designs.
- When using randomised designs it is important to consider the appropriate use of cluster, rather than individual, randomisation. This has implications for both study design and analysis.
- Where randomised designs are not feasible, non-randomised designs can be used although they are more susceptible to bias.

deductive process in which the investigator has to systematically think through how each of the internal validity threats may have influenced the data. Then the investigator has to examine the data to test which relevant threats can be ruled out. . . . When all of the threats can plausibly be eliminated it is possible to make confident conclusions about whether a relationship is probably causal." Within quasi experiments there are potentially greater threats to internal validity and less ability to account for these. We believe that the design and conduct of quasi-experimental studies is at least as methodologically challenging as the design and conduct of randomised trials. Furthermore, there has been a lack of development of quasi-experimental methods since Cook and Campbell published their key text "Quasi-experimentation: design and analysis issues for field settings" in 1979.[27] The generalisability of quasi-experimental designs is also uncertain. Many quasi-experimental studies are conducted in a small number of study sites which may not be representative of the population to which the researcher wishes to generalise.

CONCLUSIONS

We have considered a range of research designs for studies evaluating the effectiveness of change and improvement strategies. The design chosen will reflect both the needs (and resources) in any particular circumstances and also the purpose of the evaluation. The general principle underlying the choice of evaluative design is, however, simple—those conducting such evaluations should use the most robust design possible to minimise bias and maximise generalisability.

ACKNOWLEDGEMENTS

The Health Services Research Unit is funded by the Chief Scientist Office, Scottish Executive Department of Health. The views expressed are those of the authors and not the funding bodies.

Authors' affiliations

M Eccles, Centre for Health Services Research, University of Newcastle upon Tyne, Newcastle upon Tyne, UK
J Grimshaw, Clinical Epidemiology Programme, Ottawa Health Research Institute, Ottawa, Canada
M Campbell, C Ramsay, Health Services Research Unit, University of Aberdeen, Aberdeen, UK

REFERENCES

1 **Russell IT**. The evaluation of a computerised tomography: a review of research methods. In: Culyer AJ, Horisberger B, eds. *Economic and medical evaluation of health care technologies*. Berlin: Springer-Verlag, 1983:38–68.
2 **Campbell DT**, Stanley J. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally, 1966.
3 **Cook TD**, Campbell DT. *Quasi-experimentation: design and analysis issues for field settings*. Chicago: Rand McNally, 1979.
4 **Campbell M**, Fitzpatrick R, Haines A, *et al*. Framework for design and evaluation of complex interventions to improve health. *BMJ* 2000;**321**:694–6.
5 **Cochrane AL**. *Effectiveness and efficiency: random reflections on health services*. London: Nuffield Provincial Hospitals Trust, 1979.
6 **Pocock SJ**. *Clinical trials: a practical approach*. New York: Wiley, 1983.
7 **Morgan M**, Studney DR, Barnett GO, *et al*. Computerized concurrent review of prenatal care. *Qual Rev Bull* 1978;**4**:33–6.
8 **Donner A**, Klar N. *Design and analysis of cluster randomization trials in health research*. London: Arnold, 2000.
9 **Murray DM**. *The design and analysis of group randomised trials*. Oxford: Oxford University Press, 1998.
10 **Donner A**, Koval JJ. The estimation of intraclass correlation in the analysis of family data. *Biometrics* 1980;**36**:19–25.
11 **Diwan VK**, Eriksson B, Sterky G, *et al*. Randomization by group in the studying the effect of drug information in primary care. *Int J Epidemiol* 1992;**21**:124–30.
12 **Flynn TN**, Whitley E, Peters TJ. Recruitment strategies in a cluster randomised trial - cost implications. *Stat Med* 2002;**21**:397–405.
13 **Donner A**. Some aspects of the design and analysis of cluster randomization trials. *Appl Stat* 1998;**47**:95–113.
14 **Turner MJ**, Flannelly GM, Wingfield M, *et al*. The miscarriage clinic: an audit of the first year. *Br J Obstet Gynaecol* 1991;**98**:306–8.
15 **Campbell MK**, Mollison J, Steen N, *et al*. Analysis of cluster randomized trails in primary care: a practical approach. *Fam Pract* 2000;**17**:192–6.
16 **Mollison JA**, Simpson JA, Campbell MK, *et al*. Comparison of analytical methods for cluster randomised trials: an example from a primary care setting. *J Epidemiol Biostat* 2000;**5**:339–48.
17 **Lobo CM**, Frijling BD, Hulscher MEJL, *et al*. Improving quality of organising cardiovascular preventive care in general practice by outreach visitors: a randomised controlled trial. *Prevent Med* 2003 (in press).
18 **Cochran WG**, Cox GM. *Experimental design*. New York: Wiley, 1957.
19 **Eccles M**, Steen N, Grimshaw J, *et al*. Effect of audit and feedback, and reminder messages on primary-care radiology referrals: a randomised trial. *Lancet* 2001;**357**:1406–9.
20 **Eccles M**, McColl E, Steen N, *et al*. A cluster randomised controlled trial of computerised evidence based guidelines for angina and asthma in primary care. *BMJ* 2002;**325**:941–7.
21 **Lipsey MW**, Wilson DB. The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. *Am Psychol* 1993;**48**:1181–209.
22 **Matowe L**, Ramsay C, Grimshaw JM, *et al*. Influence of the Royal College of Radiologists' guidelines on referrals from general practice: a time series analysis. *Clin Radiol* 2002;**57**:575–8.
23 **Ostrom CW**. *Time series analysis: regression techniques*. London: Sage, 1990.
24 **Grilli R**, Ramsay CR, Minozi S. Mass media interventions: effects on health services utilisation In: Cochrane Collaboration. *The Cochrane Library*. Oxford: Update Software, 2002.
25 **Grilli R**, Freemantle N, Minozzi S, *et al*. Impact of mass media on health services utilisation. In: Cochrane Collaboration. *The Cochrane Library*. Issue 3. Oxford: Update Software, 1998.
26 **Schwartz D**, Lellouch J. Explanatory and pragmatic attitudes in clinical trials. *J Chron Dis* 1967;**20**:648.
27 **Shaddish WR**. The empirical program of quasi-expermentation. In Bickman, ed. *Research design*. Thousand Oaks: Sage, 2000.