Experimental and quasi-experimental designs for evaluating guideline implementation strategies

Jeremy Grimshaw, Marion Campbell, Martin Eccles[a] and Nick Steen[a]

Grimshaw J, Campbell M, Eccles M and Steen N. Experimental and quasi-experimental designs for evaluating guideline implementation strategies. *Family Practice* 2000; **17:** S11–S18.

The choice of study design for guideline implementation studies will determine the confidence with which the observed effects can be attributed to the interventions under study. In general, cluster randomized trials, of which there are different types, provide the most robust design. However, the use of these designs has implications for the power, conduct and analysis of studies. Wherever possible, designs allowing head-to-head comparisons, which incorporate baseline measures of performance, should be used.

Introduction

If policy makers are to make evidence-based decisions about guideline implementation, they need reliable information on the effectiveness and cost-effectiveness of different interventions (in different settings for different targeted clinicians and behaviours), the likely effect modifiers and the resources needed to deliver interventions. In order to obtain such information, policy makers need to use the evidence from studies that adopted rigorous designs and methods. Many existing studies use weak designs or are methodologically flawed with potentially major threats to validity, thereby limiting their value to inform decision-making.[1] In this paper, we describe a number of potential study designs, highlighting their advantages and disadvantages. Other important aspects of design are summarized elsewhere[2] and in the other papers in this supplement.

A variety of study designs can be used to evaluate guideline implementation strategies. These vary in the degree to which they allow observed effects to be attributed to the intervention with confidence.

Observational studies

Observational (or descriptive) studies of single groups may usefully provide greater understanding of the process of behavioural change and generate hypotheses for further testing in rigorous evaluations.[3] However, they are rarely useful for evaluation because the characteristics of the populations to be compared may differ in ways that affect the outcomes being measured—characteristics other than the interventions to be compared. If the evaluator cannot identify or measure these differences, nothing can be done to ameliorate the resulting bias. Even when it is possible to adjust for recognized differences, it is never possible to rule out unrecognized bias with confidence.

Quasi-experimental designs

Quasi-experimental studies often are conducted where there are practical and ethical barriers to conducting randomized controlled trials. In this section, we discuss the three most commonly used designs in guideline implementation studies: (i) uncontrolled before and after studies; (ii) time series designs; and (iii) controlled before and after studies. However, there are many different possible designs, and the reader should refer to Cook and Campbell[4] for further discussion of quasi-experimental studies.

Uncontrolled before and after studies
Uncontrolled before and after studies measure provider performance before and after the introduction of an intervention (e.g. dissemination of guidelines) in the same study site(s) and any observed differences in performance are assumed to be due to the intervention. Uncontrolled before and after studies are relatively simple to conduct and are superior to observational studies; however, they are intrinsically weak evaluative designs,[5] as secular trends or sudden changes make it difficult to attribute

Health Services Research Unit, University of Aberdeen, Aberdeen AB25 2ZD and [a]Centre for Health Services Research, University of Newcastle upon Tyne, Newcastle upon Tyne NE2 4AA, UK.

observed changes to the intervention.[4] Furthermore, in such studies, the intervention is confounded by the Hawthorne effect (the non-specific beneficial effect on performance of taking part in research)[6] which could lead to an overestimate of the effectiveness of an intervention.

There is also some evidence to suggest that the results of uncontrolled before and after studies may over-estimate the effects of interventions. Lipsey and Wilson undertook an overview of meta-analyses of psychological, educational and behavioural interventions.[7] They identified 45 reviews that reported separately the pooled estimates from controlled and uncontrolled studies; 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 guideline implementation strategies, and the results of studies using such designs have to be interpreted with great caution.

Time series designs
Time series designs attempt to detect whether an intervention has had an effect significantly greater than the underlying trend.[4] They are useful in guideline implementation research for evaluating the effects of interventions when it is difficult to randomize or identify an appropriate control group (e.g. following the dissemination of national guidelines or mass media campaigns). Data are collected at multiple time points before and after the intervention; the multiple time points before the intervention allow the underlying trend to be estimated, the multiple time points after the intervention allow the intervention effect to be estimated accounting for the underlying trend (Fig. 1). A number of statistical techniques can be used depending on the characteristics of the data;[4] the most important determinant of technique is the number of data points prior to the intervention to provide a stable estimate of the underlying trend. As a rule of thumb, 20 data points are needed before and 20 data points after the intervention to allow full time series modelling to be used.[8]

Time series designs 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. Currently, many published interrupted time series have been analysed inappropriately, frequently overestimating the effect of the intervention.[9]

Controlled before and after studies
In controlled before and after studies, a control population is identified which has similar characteristics and performance to the study population and is expected to experience secular trends or sudden changes similar to the study population.[4,10] Data are collected in both populations contemporaneously using similar methods before and after the intervention is introduced in the study population. A 'between group' analysis comparing performance in the study and control groups following the intervention is undertaken, and any observed differences are assumed to be due to intervention (Fig. 2).

Whilst 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 is often observed to differ. Under these circumstances, 'within group' analyses (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) are often undertaken. Such analyses are inappropriate for a number of reasons. Firstly, the
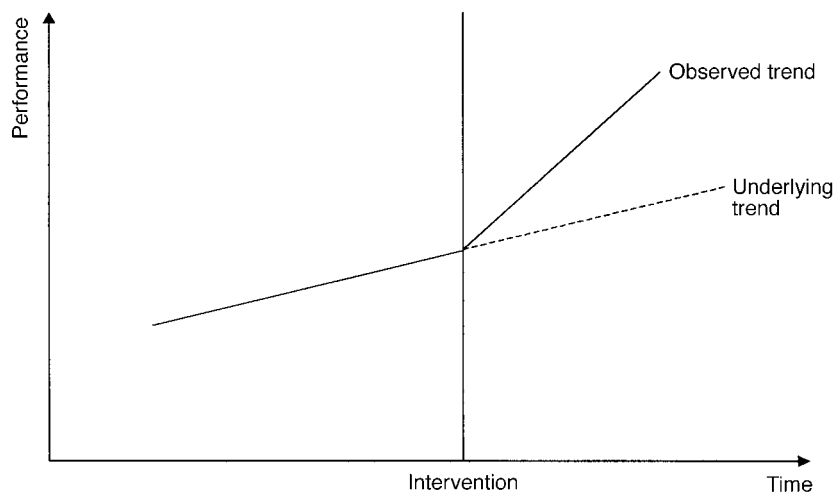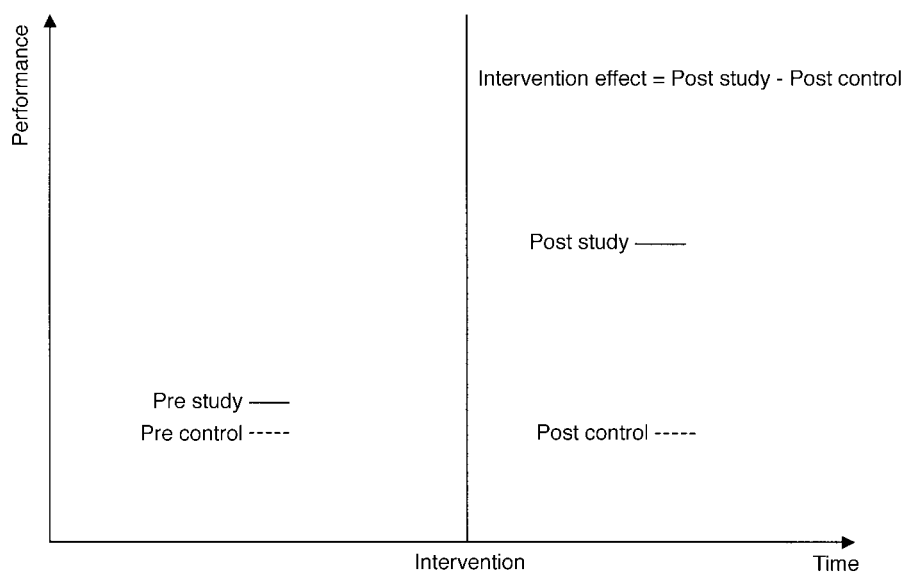
FIGURE 1   *Time series analysis*

FIGURE 2   *Controlled before and after study*

baseline imbalance suggests that the control group is not 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.

The usefulness of controlled before and after studies is also limited because the estimate of effect cannot be attributed to the intervention with confidence due to the non-randomized control group. In many circumstances, where a controlled before and after design is proposed, a randomized trial could be undertaken just as easily and would provide a more reliable estimate of effect.

Randomized trials

Patient randomized trials
Randomized trials are rightly considered to be the most robust method of assessing health care innovations.[11] Randomized trials estimate the impact of an intervention through direct comparison with a randomly allocated control group that receives either no intervention or an alternative intervention.[12] The randomization process ensures that, all else being equal, both known and unknown biases are distributed evenly between the trial groups.

When evaluating guideline implementation strategies, however, simple (patient) randomized trials may be less robust. There is a danger that the treatment offered to control patients will be contaminated by doctors' experiences of applying the intervention to patients receiving the experimental management, with the result that the evaluation may underestimate the true effects of strategies. For example, Morgan and colleagues

undertook a study of computerized reminders for antenatal care.[13] They chose to randomize patients between a control group and an experimental group for whom any non-compliance by the doctor generated an automatic reminder from the computer-based medical record system. Compliance in experimental patients rose from 83 to 98% within 6 months, while compliance in control patients rose from 83 to 94% in 12 months. The results suggest that the intervention had a significant (if delayed) effect on the management of control patients.

Cluster randomized trials
To overcome these problems, it is possible to randomize (groups of) professionals rather than individual patients. In such circumstances, data are collected about the process and outcome of care at the individual patient level. Such trials, which randomize at one level and collect data from a different level, are known as cluster randomized trials.[14,15]

Whilst cluster randomization overcomes to a large extent the problem of contamination in patient-randomized trials, it has implications for the planning, conduct and analysis of studies. A fundamental assumption of the patient-randomized trial is that the outcome for an individual patient is completely unrelated to that for any other patient, i.e. they are said to be 'independent'. This assumption is violated, however, when cluster randomization is adopted, because patients within any one cluster are more likely to respond in a similar manner. For example, the management of patients in a single hospital is more likely to be consistent than management across a number of hospitals. The primary consequence of adopting a cluster randomized design is that it is not as statistically efficient and possesses lower statistical power than a patient-randomized trial of

equivalent size (see Donner[16] and Bland [in this supplement][17] for further details). Because of this lack of independence, sample sizes require to be inflated to adjust for the clustering effect,[17] and special analytical techniques, such as multi-level modelling[18] need to be adopted, unless simple cluster-level analysis is undertaken.

Despite the added complexity, cluster randomized trials provide the optimal design for guideline implementation studies. In the next section, the advantages and disadvantages of different approaches to the conduct of cluster randomized trials are discussed.

Possible types of cluster randomized trials

Two-arm trials

The simplest design is the two-arm trial where groups of professionals are randomized to study or control groups. Such trials are relatively straightforward to design and operationalize, and they maximize power (half the sample are allocated to the intervention and half to the control). If two-arm trials are used to evaluate a single intervention against control, however, they only provide limited information about the effectiveness of the intervention within a single setting. They do not provide information about the relative effectiveness of different interventions within the same setting. Two-arm trials can also be used to compare two different interventions, but do not provide information about the effect of either intervention against a control.

Because the simple two-arm trial does not provide data on the relative effectiveness of different interventions, they are of limited value to decision-makers. For this reason, results from extensions to the two-arm trial such as the multi-arm trial or the factorial design may prove more informative.

Multiple arm trials

The simplest extension to the two-arm trial is to randomize groups of professionals to more than two groups (e.g. 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 different 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 comparison of more than one intervention with reduced loss of power compared with multiple arm trials. For example, in a 2×2 factorial design evaluating two interventions against control, participants are randomized to each intervention (A and B) independently (see Table 1). In the first randomization,

the study participants are randomized to intervention A or control. In the second randomization, the same participants are randomized to intervention B or control. This results in four groups: no intervention, intervention A only, intervention B only, and both intervention A and B. During the analysis of factorial designs, it is possible to undertake independent analyses to estimate the effect of the interventions separately;[19] essentially, this design allows two randomized trials to be conducted for the same sample size as a two-arm trial. However, these trials are more difficult to operationalize 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.

Balanced incomplete block designs

In guideline implementation research, there are also a number of non-specific effects, which may influence the estimate of the effect of an intervention. Currently, these non-specific effects are lumped together and termed the 'Hawthorne effect'. If these are imbalanced across study groups in guideline implementation trials, the resulting estimates of effects may be biased.

Balanced incomplete block designs can be used to equalize such non-specific effects and thereby minimize their impact.[19] The simplest design is a 2×2 balanced incomplete block design (Table 2). The study population is allocated randomly between two groups. One group receives the intervention for the management of condition one and provides control data for the management

TABLE 1 *Diagrammatic representation of a factorial design*

Randomization to intervention B	Randomization to Intervention A	
	No	Yes
No	Group 1 receive neither intervention	Group 2 receive intervention A only
Yes	Group 3 receive intervention B only	Group 4 receive both interventions

TABLE 2 *Example of balanced incomplete block design*

	Condition 1	Condition 2
Study group A	Intervention	Control
Study group B	Control	Intervention

of condition two. The other group receives the intervention for condition two and provides control data for condition one. As subjects in both groups experience the same level of intervention, the Hawthorne effect should be equalized across the two groups. Such designs should enhance the generalizability of the study findings as they test the effects of the intervention across different conditions. However, they are complex to design, operationalize (especially if more complicated balanced incomplete block designs are used to test different levels of intervention or different interventions) and analyse.

Baseline measurement in cluster randomized trials

Commonly in cluster randomized trials, relatively few clusters (e.g. general practices) are randomized. Under these circumstances, there is increased danger of imbalance in performance between study and control groups due to chance. Baseline measurements can be used to assess the adequacy of the allocation process. Ideally, such measurement should take place in the planning or pilot stage of an implementation trial, and baseline performance should be used as a stratifying variable; this ensures balance across study and control groups with subsequent increase in statistical power. Correcting for baseline performance during analysis can also increase statistical power.[20]

In addition, baseline measures of performance are useful because they provide an estimate of the magnitude of a problem. Low performance scores prior to the intervention may indicate that performance is poor and there is much room for improvement. On the other hand, high performance scores may indicate that there is little room for improvement (ceiling effect).

Discussion

The central tenet of Professor Grimshaw's paper, that "randomized trials are rightly considered the most robust method of assessing health care innovations", was challenged by some participants. The general discussion covered a number of problems associated with randomized trials and how these might be overcome.

Randomization does not account for physician preference
It was pointed out that physicians, like all people, will only change their behaviour if they are so motivated. In ordinary practice, physicians who take up, or comply with, an intervention designed to change practice have made an explicit or implicit decision to do so. In experiments where physicians are randomized to receive (or not) a similar intervention, their motivation and attitude may differ across trial groups. This behavioural factor

may then hinder the interpretation of negative results in implementation trials. Is the intervention really ineffective, or has the artificial environment created by the trial antagonized its effect? Similarly, the lack of a positive choice in trials may explain Lipsey and Wilson's finding that uncontrolled studies showed greater effect sizes.[7]

One solution suggested was to conduct 'preference' trials.[21] However, these are difficult to perform, only a few having been reported in the literature so far. Another suggestion was to investigate, during the baseline phase of trials, the attitudes of participating clinicians towards the interventions using qualitative and psychological methods.

Randomized trials are lengthy
Decision-makers commission research to allow policy to be more evidence-based, but often the time lag before trials produce their robust results means that decisions which the results were supposed to inform have already been taken. The larger the trial, the greater the likelihood of an inconvenient delay.

There was no consensus about the best way to address this problem. One participant argued that randomized trials of interventions to change practice were being carried out far too early in their life cycle. Pharmaceutical products are only trialled on a large scale once a number of other types of studies have been completed and there is a fair expectation that the drug will become part of established practice. By contrast, randomized trials of behavioural interventions might be proposed as the first stage of evaluation.

Encouragement of small-scale observational studies and anecdotal reports was an alternative approach. The recently established NHSnet 'Learning Zone' was an example. Some participants were concerned that this would encourage inappropriate service developments, particularly if marks of approval, such as 'beacon' status, were awarded to unevaluated schemes.

A 'third way' suggested was to undertake a stream of small randomized studies, and from these pick one or two promising interventions for full-scale trials. A modification of this approach, suggested by Professor Grimshaw, was to pick specific behavioural issues and investigate them qualitatively at the same time as undertaking a trial. One advantage would be that several disciplines could contribute to the research.

The number of subjects required for statistical power
As pointed out above, many trials have lacked adequate power, particularly for the analyses that matter most to policy makers, such as head-to-head comparisons of alternative interventions or interactions with baseline characteristics.

The solution to date had been to develop networks of clinicians who were willing to act as subjects. In the UK, several such networks exist in primary care, but little progress has been made in hospitals. There are problems,

however, with this approach: such networks are not necessarily representative; each tends to be developed by a particular researcher; and consent for participation in trials is generally a difficult ethical issue. (Winkens and colleagues have reported an approach for conducting intervention trials without obtaining study-specific informed consent.[22])

Acknowledgements

References

1   Bero L, Grilli R, Grimshaw JM, Harvey E, Oxman AD, Thomson MA. Closing the gap between research and practice: an overview of systematic reviews of interventions to promote implementation of research findings by health care professionals. *Br Med J* 1998; **317:** 465–468.

2   Campbell MK, Steen IN, Grimshaw JM, Eccles MP, Mollison JA, Lombard C. Design and statistical issues in implementation research. In Makaela M, Thoreson T (eds). *Changing Professional Practice*. Copenhagen: Danish Institute of Health Services Research and Development, 1999; 57–76.

3   Grilli R, Lomas J. Evaluating the message: the relationship between compliance rate and the subject of practice guideline. *Med Care* 1994; **32:** 202–213.

4   Cook TD, Campbell DT. *Quasi-experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally, 1979.

5   Russell IT, Grimshaw JM. The effectiveness of referral guidelines: a review of the methods and findings of published evaluations. In Coulter A and Roland M (eds). *Referrals from General Practice*. Oxford: Oxford University Press, 1992.

6   Moser CA, Kalton G. *Survey Methods in Social Investigation*. 2nd edn. Aldershot: Gower, 1979.

7   Lipsey MW, Wilson DB. The efficacy of psychological, educational and behavioural treatment. Confirmation from meta-analysis. *Am Psychol* 1993; **48:** 1181–1209.

8   Crabtree BF, Ray SC, Schmidt PM, O'Connor PJ, Schmidt DD. The individual over time: time series applications in health care research. *J Clin Epidemiol* 1990; **43:** 241–260.

9   Grilli R, Freemantle N, Minozzi S, Domenighetti G, Finer D. *Impact of Mass Media on Health Services Utilisation (Cochrane Review)*. In The Cochrane Library, Issue 3 Oxford: Update Software; 1998.

10 Effective Health Care. *Implementing Clinical Guidelines. Can Guidelines be Used to Improve Clinical Practice? Bulletin No. 8*. Leeds: University of Leeds, 1994.

11 Cochrane AL. *Effectiveness and Efficiency: Random Reflections on Health Services*. London: Nuffield Provincial Hospitals Trust, 1972.

12 Pocock SJ. *Clinical Trials: A Practical Approach*. New York: Wiley, 1983.

13 Morgan M, Studney DR, Barnett GO, Winickoff RN. Computerized concurrent review of prenatal care. *Qual Rev Bull* 1978; **4:** 33–36.

14 Donner A. Some aspects of the design and analysis of cluster randomization trials. *Appl Statist* 1998; **47:** 95–113.

15 Murray DM. *The Design and Analysis of Group Randomised Trials*. Oxford: Oxford University Press, 1998.

16 Donner A, Birkett N, Buck C. Randomization by cluster, sample size requirements and analysis. *Am J Epidemiol* 1981; **114:** 906–915.

17 Bland JM. Sample size in guidelines trials. *Family Practice* 2000; **17:** Suppl 1: S17–S20.

18 Rice N, Leyland A. Multilevel models: applications to health data. *J Health Serv Res Policy* 1996; **1:** 154–164.

19 Cochran WG, Cox GM. *Experimental Design*. 2nd edn. New York: Wiley, 1979.

20 Duffy SW, South MC, Day NE. Cluster randomisation in large public health trials: the importance of antecedent data. *Statist Med* 1992; **11:** 307–316.

21 Togerson D, Sibbald B. What is a patient preference trial? *Br Med J* 1998; **316:** 360.

22 Winkens RA, Knottnerus JA, Kester AD, Grol RP, Pop P. Fitting a routine health-care activity into a randomized trial: an experiment possible without informed consent? *J Clin Epidemiol* 1997; **50:** 435–439.