THE UNIVERSITY *of York*

CENTRE FOR HEALTH ECONOMICS

Evaluating Change in Professional Behaviour: Issues in Design and Analysis

Nick Freemantle
John Wood
James Mason

DISCUSSION PAPER 171

EVALUATING CHANGE IN PROFESSIONAL BEHAVIOUR: ISSUES IN DESIGN AND ANALYSIS

Nick Freemantle[1]

John Wood[2]

James Mason[1]

1. Centre for Health Economics, University of York

2. Department of Health Sciences & Clinical Evaluation, University of York

July 1999

SUMMARY

Implementing the findings of research to change the behaviour of health care professionals has become an increasingly prominent issue. However, designing valid studies to evaluate different methods of achieving changes requires considerable care and there are a number of pitfalls evident from published previous work. The various steps in the development of an implementation method and issues arising are explored in this text. Aspects include conceptualisation, essential background work, a structured development process, the relative merits of randomised and non-equivalent group designs, the unit of analysis, the role of multi-level models, block designs, economic analysis, and the content or message to be disseminated. An ongoing, large, randomised trial of educational outreach visits by trained pharmacists is used to illustrate some of the issues.

Keywords
Behavioural change, implementation methods, economic evaluation, design of trials.

INTRODUCTION

The recent National Health Service White paper places great emphasis on quality assurance, one facet being clinicians' awareness and application of the best evidence available for the treatment of patients. [1] However, it is apparent that important findings from large randomised trials and systematic overviews in areas such as the treatment of heart failure, [2,3] antiplatelet therapy [4,5] and thrombolytic therapy [6,7] do not translate automatically into practice, [8-10]. The process of trying to implement research findings raises a number of difficulties. [11,12] As David Eddy has commented, while the medical profession has placed a high value on developing the basic science of medicine, it has not emphasized the process by which that science is translated into practice. [12]

The English NHS has recognised explicitly the need to focus effort on the implementation issue [13] and has recently commenced a programme of research focusing largely on randomised and quasi-experimental studies of interventions to change practice. [14] This text explores the design and analysis issues for methods intended to promote appropriate behavioural change. Some of the material is technical and assumes knowledge of the design of studies to evaluate health care treatments. Building from this starting point, the intention is to aid clarity of thought on the part of those involved in designing or evaluating implementation research.

First, the importance of undertaking initial qualitative work is discussed. Second, we review the characteristics of experimental and quasi-experimental studies and the interpretation of their results. Third, we consider issues that impact upon the choice of the unit of analysis in evaluations of interventions intended to change practice. Fourth, we explore block study designs as a mechanism for dealing with problems of variation and bias peculiar to implementation studies. Finally, we examine the design issues for the inclusion of economics.

ASKING THE RIGHT QUESTIONS

Doing the developmental work first

It is fundamental to any research to begin by asking the right questions and then choosing the right methods to answer them. This may seem obvious, but it is possible to identify quantitative research that might have been considered unnecessary or poorly focused had qualitative work been used to judge whether the hypotheses under test were sensible. For example, the perceived need for information by primary care physicians was identified in one small trial assessing intermediate outcome measures. [15] Its provision was subsequently evaluated within a larger randomised trial with objective measurement of patient and health professional outcomes. [16] The outcome of the larger trial was that the simple provision of information to doctors did not appear to bring substantial direct benefits to patients, a conclusion confirmed in a subsequent systematic review of this question. [10] In an observational study, Covell and colleagues [17] identified that the perceived need for better information among office-based practitioners was not reflected in their behaviour. Although doctors perceived that they answered questions arising from their work through traditional information sources such as textbooks and journal articles, they actually resolved such issues through consultation with colleagues. Thus, it was, *a priori,* unlikely that the provision of information by itself would be effective. In this instance, the quantitative evaluation was ill-focused when considered in the light of the relevant qualitative research.

A sequence of questions

Historically, most randomised trials in the medical field have been conducted with drugs, [18] reflecting both the dominance of this form of treatment in modern western health care systems and the financial and regulatory state of the pharmaceutical industry. Those attempting to develop and evaluate interventions to increase the uptake of research findings might usefully reflect on the processes in the development of drugs. Research and development in the pharmaceutical field commences a very long way from patients, and also a long way from clinical experiments, with substantial pre-clinical work aimed at developing and describing the attributes of new chemical entities. Having sorted out how such entities behave in test tubes and animals, their action is then examined in healthy volunteers, and further evaluated in dose ranging studies. Only eventually are double blind randomised trials conducted in groups of diseased individuals. [18] Stages of pharmaceutical development and corresponding stages in the development of implementation interventions are described in Table 1. All the diverse and extensive pre-clinical work has been lumped together and called ‘Phase 0’. In the context of the development of interventions intended to influence practice, this important stage should include qualitative work to describe and develop an understanding of the effects of an implementation approach.

Interventions aiming to introduce research findings to practice, such as computerised decision support systems or audit and feedback, require careful development, although most have not been through such a development process. Educational outreach (academic detailing) aiming to change prescribing practice is probably the exception, with substantial development of the theory and practice of the approach [19] followed ultimately by the evaluation of its efficiency in different health systems. [20] Where interventions are not well established and described, investigators risk evaluating the ‘wrong kind’ of audit and feedback or reminder system. Evaluative trials of inadequately developed and understood implementation interventions are premature.

Explanatory or pragmatic design?

An explanatory design aims to increase understanding of the intervention under study, while a pragmatic design addresses the size of effect in regular practice. These different aims have important implications for design and analysis. The distinction can also be used to help crystallise the purpose of an evaluative trial beyond the ill-specified wish to ‘test out’ the treatments of interest, as the two types of experiment have conflicting strengths.

Consistent with the natural sequence of development shown in Table 1, it is important to establish that an intervention *can* work and to gain an understanding of the *way* it works in studies with high construct validity *before* attempting to evaluate whether it can achieve this potential in the real world.

Table 1: Phases of clinical trials and proposed comparable phases for implementation evaluations.

Phase	Drug Trial	Implementation Method
Phase 0	Premedical research containing many stages, including studies *in vitro* and in animals, aiming to describe the action of a drug in different (artificial) circumstances.	Developing new ideas and using qualitative research to examine their effects (e.g. developing and qualitatively evaluating a computer decision support system)
Phase 1	Studies in healthy volunteers, aiming to establish safety rather than efficacy of a drug.	Testing the applicability of an intervention in artificial but demanding circumstances (e.g. testing a computerised decision support system with a classroom of computer science students for a day).
Phase 2	Small-scale investigations of treatment efficacy in diseased individuals, including dose ranging, often aiming to screen out chemical entities not likely to be clinically useful.	Examining the feasibility of an implementation intervention in practice (e.g. assessing the potential of a computerised decision support system in clinical practice in terms of comprehension of prompts and reaction/satisfaction of GPs).
Phase 3	Comparison with current standard treatment in large-scale rigorous experimental studies in diseased individuals.	Large-scale trials examining the effectiveness of an intervention in controlled experimental situations (e.g. testing whether the addition of computerised decision support has the desired effect on practice against mailed information alone).
Phase 4	Studies of the use of a drug in established practice (post-marketing surveillance/ marketing). Focus on rare side effects and potential impact of drug on practice.	Establishing the effectiveness and efficiency of implementation techniques in real world settings (e.g. assessing whether widespread implementation of promising computerised decision support systems has had the desired impact upon practice).

For example, consider the design of a trial to evaluate a computer prompt system providing treatment guidance for a certain clinical condition. In addition, suppose the guidance to be provided is generally available in the clinical literature. A pragmatically designed trial would use a real-world comparison of no formal intervention (clinicians may or may not be aware of the literature). The computer hardware and software would be provided and operated under 'normal' conditions (through a market provider) and no incentives would be offered to clinicians to participate (unless these are present under normal service conditions). An explanatory design would feature 'equalised' conditions: the control group would all be presented with the guidance on paper, the computer prompt system would be provided and monitored by the trialists, and incentives may be provided to ensure compliance with the protocol. Both designs explore the extent to which intervention changes behaviour, although their findings relate to real world and optimised settings.

The pragmatic design, while clearly more applicable to decision-makers attempting to set policy, is vulnerable to a number of confounding influences. A negative pragmatic trial result may reflect the fact that the computer prompt system was inadequately designed, that the computer supplier particular to the trial provided inadequate service support, or that the clinicians simply ignored the prompts. A positive result could include some behavioural modification through improving participants' awareness of the main study outcomes; in other words, it may not be specific to computerised decision support. There is seldom an absolute dichotomy between 'pragmatic' and 'explanatory' but in practice a continuum and these issues should be addressed at the design stage. For example, where distribution of educational materials is the norm, there may be no difference between explanatory and pragmatic design with respect to the choice of the control intervention.

Experimental and quasi-experimental studies

There are two main forms of experimental designs that may provide valid estimates of the effect of interventions - randomised trials and non-equivalent group designs. [21] In randomised trials, subjects are allocated by chance to either treatment or control groups, and so bias is distributed between the groups by chance. In other words, treatment and control groups may be considered equivalent, apart from the play of chance. Non-equivalent group study designs allocate subjects without random allocation, and thus differences between the groups may reflect systematic biases.

Advantages of randomisation

Randomised trials are rightly regarded as the most valid means of providing a quantitative estimate of the effectiveness of health care interventions. Randomisation plays a central role in ensuring that any extraneous factors (both known and unknown) have an equal chance of affecting each treatment group. However, it does not ensure that each treatment group will be comparable, as allocation depends upon the play of chance. Results are also expected to differ according to chance but the process of randomisation ensures that we have a valid estimate of the expected size of this variation (through the estimate of experimental error). Thus randomisation can be said to ensure the validity of the statistical analysis. In situations where randomisation is practical, there seems no good reason for omitting it. Unfortunately, after the randomisation has been carried out, the groups have an opportunity to diverge, simply as a consequence of being labelled differently, rather than as a result of a specific treatment effect. In drug trials, this problem is addressed by making them double-blind which, if done successfully, should be a complete answer. For professional behaviour-change trials, blinding is not an option and labelling effects need to be addressed in a different way.

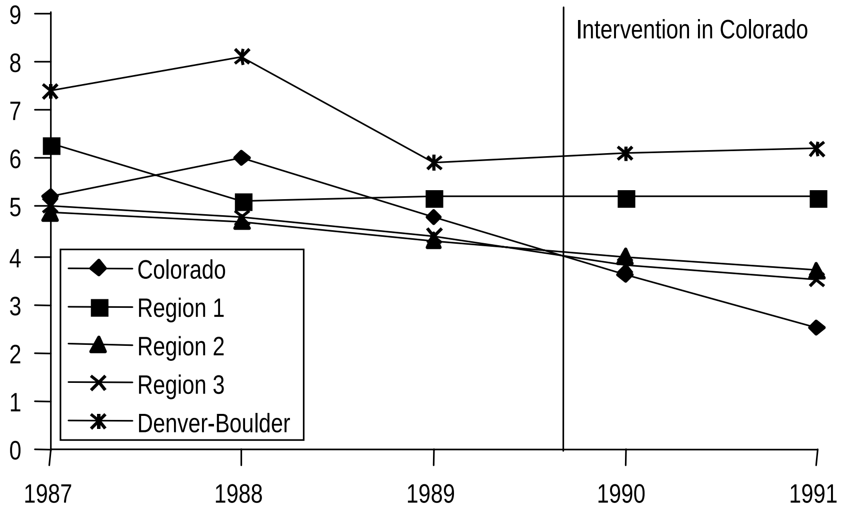
Non-equivalent group design studies

In certain circumstances randomisation is not possible, such as national guideline implementation programmes or mass media campaigns. Although potentially useful in these circumstances, non-equivalent group designs require more careful interpretation than randomised studies. There are three main potential biases that may affect such studies. First, there may be scaling differences in which apparent differences in effect sizes between groups over time are in fact associated with the instrument of measurement rather than real differences in effect. Second, it may be that different groups are at different stages in a common maturational process, and thus respond to interventions in ways that may exaggerate or mask the true treatment effect. One group may be merely lagging behind another in time, and an effect - either beneficial or otherwise - apparently associated with an intervention, may be spurious. Third, the likelihood of chance differences may be high, and groups may, at any time, be at different positions on a common (often cyclical) process. Initial outliers in groups may be expected over time to regress to the mean, and apparent differences may simply reflect this underlying chance process. The second and third kinds of bias might be considered special cases of selection bias in general, which in a range of guises makes the attribution of apparent responses to treatments or interventions problematic. [21]

One partial solution to the problem of attribution caused by non-equivalence is the collection of additional data points at time periods before and after the intervention. [21] A study in which there are only post intervention data points is potentially very misleading, as any contribution resulting from the intervention cannot be disentangled from what might have happened anyway. Adding extra data points before and after the intervention enables some estimation of concurrent trends and the contribution, if any, of the intervention.

The usefulness of additional data points in interpreting non-equivalent group designs is demonstrated by data from the study by Wagner et al (1995) [22]. The study describes the impact achieved on rates of surgery for benign prostatic hyperplasia in US patients by a video disk programme to inform patients of treatment options, (Figure 1). Patients in the intervention group, Colorado, received the intervention from September 1989. All other districts described in the figure served as controls. Had data been available only from September 1989 onwards, we would have been unaware that the rate of change in surgery for Colorado was actually unchanged from the year before the intervention began, and that the overall rate of change for all groups was more similar than different over the study period as a whole.

Figure 1: Surgery for Benign Prostatic Hyperplasia

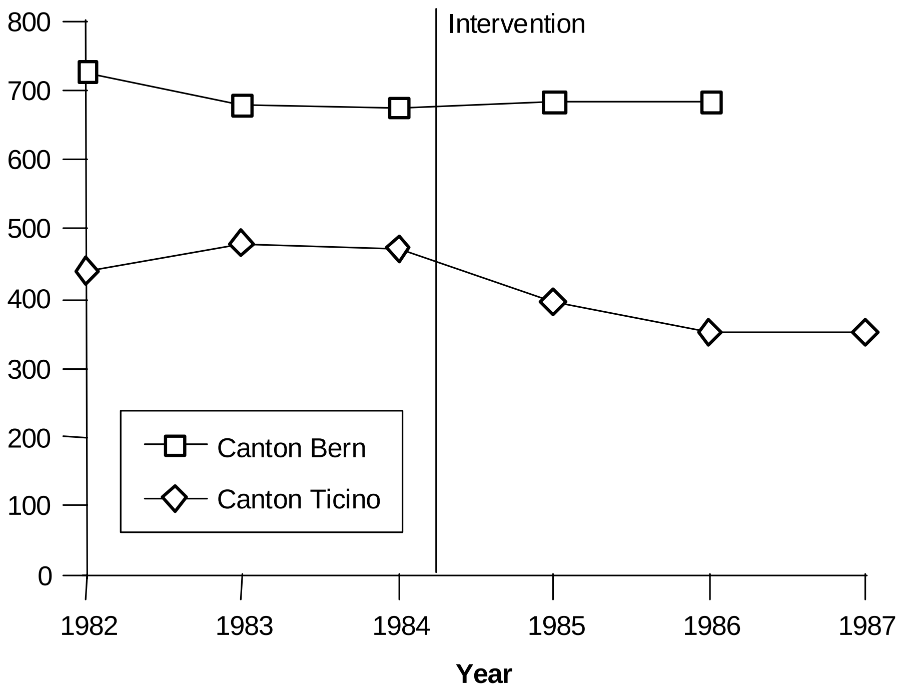


Transurethral prostatectomy rates in intervention and control regions / 1000 men 45 years or older. Adapted from Wagner 1995[27]

Another potential challenge to the validity of non-randomised, quasi-experimental studies is the contexturalising of results, limiting their generalisability to other settings. The underlying principle of experimentation is that the treatment is introduced by the experimenter and the impact, or otherwise, is assessed. Where treatments are allocated randomly, there is a strong likelihood that the treatment may be separated from its context. Where treatments are allocated in other ways (non randomly) this may not be the case. Even in randomised trials, selection bias has, for some time, been acknowledged as potentially important both for those receiving and providing treatment. [23] Just as practitioners may be reluctant to treat patients 'on the toss of a coin', they may also be reluctant to receive interventions intended to help them provide more effective and efficient care on this same basis.

In the example of surgery for benign prostatic hyperplasia above (Figure 1), the introduction of video disks may reflect existing concerns about the rate of intervention - a symptom of, rather than a cure for, the situation. Another study of the impact of a mass media campaign explored rates of hysterectomy in the Ticino Canton in Switzerland. [24] The campaign was concerned with the inappropriateness of high levels of intervention and appears to have arisen spontaneously through a form of *media advocacy* [25], rather than having been induced by the investigators. The comparator, Canton Bern, was clearly not equivalent, at baseline, in its use of hysterectomy. Although the annual rate of hysterectomy decreased after the mass media campaign in Ticino, a small decrease during the year before the intervention may indicate the beginning of a cyclical change (see Figure 2).

Figure 2 Hysterectomy rates per 100,000 women in 2 Swiss Cantons



The neighbouring Canton, Bern, offered as a control, had experienced a fairly sharp change in rates between 1981 and 1982 for which no explanation is given. More fundamentally, concern among clinicians, leading to a change, may have been sparked by an article by a chief surgeon of a public hospital which itself was responsible for motivating the mass media campaign. Thus, without random allocation, it is not possible to attribute cause and effect and the investigators' statement that 'there seems to be no doubt that the decrease in hysterectomies in Ticino was the result of the mass media information campaign', appears overstated.

THE UNIT OF ANALYSIS IN IMPLEMENTATION STUDIES

In general, there is a natural structure to an experiment that exists quite apart from any treatment-related differences we may impose. Often, this is in the form of a hierarchical framework, where patients are grouped together according to their family doctor, the family doctors by practice and the practices by health authority. Naturally, this has consequences for the similarities and differences we expect to see. Patients of the same family doctor are likely to receive more similar treatment than patients under different family doctors: this should be reflected in the design and analysis of any experiment we choose to do.

In many medical trials, there is an obvious 'principal' unit - the patient. The intervention is often easy to deliver reproducibly (a particular drug, say) and aimed directly at the patient, whose progress constitutes the outcome of the experiment. Additionally, patient to patient variation is usually the major extraneous source of 'experimental error'. Here we will naturally analyse at the level of the patient and, in the absence of compelling reasons to the contrary, we shall want to randomise at the level of the patient also. In multi-centre trials, we may well find treatment effects varying from centre to centre, but as a rule, we shall seek to explain these centre-treatment interactions, rather than using them as estimates of error with a new unit of analysis. However, for trials of interventions attempting to influence practice, there is often quite a 'deep' hierarchical structure with no obvious 'principal unit'. It is this complexity of structure that can cause difficulty in the choice of the unit of analysis.

Replication and randomisation

As discussed in the last section, randomisation distributes potential confounding factors by play of chance and validates the assumptions upon which the usual statistical analysis is based. Replication in an experiment serves to increase the precision of the estimates of treatment effects by sheer weight of numbers. Where treatment effects are expected to be small and the influence of other factors relatively large, substantial replication is required to overcome the background variation. This is why, for instance, investigators elected to randomise over 40,000 patients in the GUSTO trial, which compared four different therapeutic regimens for acute myocardial infarction. [26] Because active treatment regimens were being compared and relatively small comparative effects estimated, substantial replication was required to achieve adequate levels of precision.

Of course, in the GUSTO trial, the patients were randomised individually to the treatments under study, but in experiments with a hierarchical structure, such individual allocation may not be practicable (or even possible) and instead, treatments may be randomised to groups within the hierarchy. Such allocation of groups of subjects rather than individuals is sometimes called ‘cluster randomisation’ and it has long been recognised that analysing such trials as if individuals had been allocated separately may be quite wrong. [27]

Selecting an appropriate unit of analysis

Naturally, we want a trial to provide a good estimate of the underlying population value for the outcome of interest and in order to achieve a sensible interpretation of the results, we need a measurement of the uncertainty associated with that estimate. The effect that the choice of unit of analysis has on our measurement of uncertainty (or ‘error’) becomes clear if we examine how we arrive at the latter.

The error estimate comes from the data, using the differences in outcome between units that received the same treatment. The number of such independent comparisons is termed the ‘residual degrees of freedom’. Pooling the information from these gives an estimate of the background variance and dividing this by the replication of each treatment (degrees of freedom) gives estimates of the variances of the treatment means. The square roots of these - the standard errors - can then be used to construct confidence intervals for the treatment effects. When the outcome is binary - e.g. survival after six months - the error estimates are constructed from a theoretical model - the binomial distribution - but they can be thought of in the same way.

In other words, the choice of unit of analysis has a critical effect upon the interpretation of the results of an experiment, as it determines the number of degrees of freedom on which the estimation of background variance is based, affecting the estimates of standard errors - and hence the statistical power - through two routes.

Where there are several options for the unit of analysis, as is often the case in trials that aim to influence physician behaviour, the right approach may not be immediately apparent. There is no magic formula through which the correct unit of analysis may be determined, but the answers to four questions (Table 2) should help the process.

Table 2: Questions to inform decisions on an appropriate unit of analysis

1. What is the unit of randomisation?
2. At what level is the intervention aimed (e.g. doctors, patients, hospitals)?
3. What is the main outcome being measured?
4. What are the major factors that affect results (other than the interventions themselves)?

The units of randomisation, intervention and outcome measure are the important levels of an experiment and the factors inducing variation at each level can differ markedly in type and importance. For the ‘simple medical experiment’ the answers to questions 2 to 4 all point to the patient being the desirable unit of analysis. If the patient is also made the unit of randomisation (question 1), this provides formal justification for an analysis on that basis.

There is always a case for making the unit of analysis the same as the unit of randomisation. Standard errors from such an analysis are valid indications of how much the results of the experiment would be expected to change had a different randomisation been used. The use of a unit of analysis different from the unit of randomisation assumes that the variation associated with the level of randomisation is unimportant.

Thus, if practices or organisational groups are randomised, one approach is to analyse the results of the experiments in terms of change *at the level of the practice or organisational group*. An intervention such as educational outreach [28] may influence not only physicians directly, but also have some sort of group effect within a practice or organisational group, this group effect itself varying between practices.

Randomising entire practices rather than doctors within practices may also be the most sensible approach, for a host of sound, practical reasons. For instance, in a study that aims to influence the prescribing behaviour of doctors within a practice, it may impractical to ascertain who actually prescribed a particular drug to a patient from computerised records or reimbursement data.

Such a randomisation scheme (i.e. by practice) would ‘fit’ well with an analysis at the practice level. However, it may be equally reasonably for the investigator to suppose that any ‘practice effect’ is small compared to the differences between doctors, as it is the doctors within practices who take decisions, rather than practices as a whole. This would suggest that it may be better to base the analysis on variability at the level of the doctor (the decision-maker) rather than the practice. Such an approach is defensible, but only if proper consideration has been given to the size of any potential group (‘cluster’) effect.

It has long been established that to ignore clustering can easily produce an artificially precise estimate of treatment effect, as the estimates of standard error then become too small, making estimates of treatment effects appear better than they are, and producing significant differences when none really exist. For example, in a comparison of process and outcome for hospitalised patients treated either by family physicians or internists, Franks and Dickson [29] observed an apparently very strong relationship between the number of diagnoses assigned and the discipline of the attending doctor. Internists assigned an average of 0.35 additional diagnoses per patient. When analysed using the 1988 patients as the unit of analysis the standard error was 0.08, and the p value < 0.0001. However, when analysed (correctly) using the 78 doctors as the unit of analysis, the standard error was 0.18 and the reported p-value 0.05.

In other words, randomisation by cluster, accompanied by an analysis appropriate to randomisation of individual patients, is an exercise in self-deception. [30] Techniques are available to take the potential “clustering” effect into account in analyses and in the estimation of sample size. [28, 31-35] These approaches calculate the effective loss of replication (and thus statistical power) as a consequence of cluster randomisation on the basis of estimates of the intra-cluster correlation coefficient, which is a measure of the variation between units, according to whether they belong to the same or to different clusters. In other words, the intra-cluster correlation coefficient describes the extent to which subjects within a cluster are truly independent of each other, and to what degree their attributes may be predicted from knowing to which cluster they belong. A correlation coefficient of zero indicates that subjects are completely independent of each other, whilst a coefficient of 1 indicates that membership of a cluster and a particular attribute are perfectly correlated.

Of course, the true intra-cluster correlation coefficient is generally unknown, and must be estimated from the available data. Unfortunately, statistical problems can arise if we need an estimate of the intra-cluster correlation coefficient when the number of clusters is small and in such cases, the consequent small number of degrees of freedom for error at that level may be inadequate for stable estimation.

When we assess the level at which an intervention is targeted, it becomes clear that for interventions aimed at influencing physicians, conducting an analysis at the level of their patients is almost always likely to be inappropriate, as it provides an inadequate representation of variability at the level of the decision-maker. In consequence, such an analysis will generally give misleadingly precise estimates of treatment effects.

While it makes sense to use patients or episodes of care to provide estimates of physician behaviour (e.g. proportion of prescriptions for drug A) and such estimates should be determined from sufficient numbers of patients to minimise measurement error, a single number applying to each physician or practice should go forward into an analysis. However, if there are unequal numbers of patients attached to each physician or practice, this may lead to errors in the unweighted point estimate of effect.

Selecting the wrong unit of analysis can be seriously misleading, but as an alternative to attempting to select a single ‘best’ level for the analysis, it is possible to consider modelling the whole hierarchy using appropriate statistical techniques.

Multi-level models / hierarchical models

Recently, work on hierarchical techniques, or multi-level models –(developed particularly by those with an interest in education [36]) has received some interest in health services research. [26,27] These approaches model the hierarchical structure of data common in health care.

Duffy and colleagues [37] demonstrate that cluster randomisation leads to a reduction in statistical power, but identifying antecedence data (or predictors of outcome) for subjects within a cluster and including this in the analysis may substantially improve the statistical power of a study. Multi-level models may make use of this phenomenon to provide estimates of the effects of an intervention at a higher level through the attributes of subjects at a lower level. In other words, the variability between doctors’ behaviour may, in part, be explained by the attributes of their patients. Multi-level modelling has been used to describe regional variations in mortality rates in England and Wales [38], though standard methods for multi-level modelling are based upon nested ordinary least squares (iterative generalised least squares) regression models, and so are equivalent to analyses based upon the estimated intra-cluster correlation coefficient, taking the estimated covariance structure to be true. This approach may be unstable where the numbers are small at any level of the hierarchical structure. [39] Indeed, Diwan and colleagues [40] in their simulation study of the implications for power in a proposed study of physicians working in health centres (which formed the clusters for analysis) described how an increase in the number of clusters, rather than an increase in the number of doctors or patients within a health centre, leads to an increase in statistical power.

The use of standard hierarchical modelling techniques suggests a somewhat paradoxical situation. The quest for additional statistical power may lead to acceptance of an approach that may only be stable in the large sample situation, yet in that situation, analysis at a higher level may already be adequate (without recourse to multi-level techniques). Fortunately, hierarchical approaches have been developed based upon techniques that adequately model sparse data, permitting analyses that take account of the potential instability of the observed covariance structure. [39-41]

Multi-level techniques have been developed with survey data very much in mind, and we are aware of no occasion when this approach has been used in an implementation trial. It is our intention to use these methods in the secondary analysis of an ongoing randomised trial of the effectiveness of educational outreach visits by community pharmacists attempting to influence prescribing in UK primary care. It should be noted that this approach may lead to a reduction in the precision in estimates of the effects between clusters when compared with an analysis based upon summary data, rather than inevitably leading to greater statistical power.
Hierarchical techniques have been used to ‘control’ for patient, diagnostic and practitioner variables in an analysis of prescribing data from New Zealand. [42] Although including patient characteristics can improve the predictive power of a statistical model, in this instance it did not explain inter-practitioner variability in prescribing rates. This conclusion is perhaps unsurprising in the context of intervention trials where it is doctors whose behaviour is being influenced. Doctors do not respond passively to variability in patients, so this may provide an inadequate model upon which to base our understanding of variations in practice, although variability in their response to different patients may in itself be of interest in health services research.

At a practical level, rather than ‘controlling’ for differences in patient characteristics, trials may prove more useful if they intervene in a representative sample of practices and examine the extent to which certain characteristics may undermine the overall results through stratified randomisation or, less desirably, in pre-specified subgroups. [43]

Some grey areas remain and in these instances it may be helpful to consider the outcome that the intervention hopes to achieve. For example, computerised decision support systems, such as those providing drug dosage advice, may be used to replace physician decision-making (and thus physician variability). Here it makes sense to analyse the experience of patients, because two different computers running the same software would be expected to react identically to the same situation. It is clear that the outcome is based at the level of the patient and the intervention used is intended to remove prescriber variability. However, where the comparison is with ‘normal practice’, there may be differences of practical importance between physicians, making analysis at the level of the physician more appropriate.

Having chosen what seems likely to be the most appropriate unit of analysis, investigators might consider examining the robustness of their results to the assumptions made through sensitivity analyses. For example, analysis at the level of the doctor rather than practice could conceivably give answers that are qualitatively different, as well as having different levels of precision. Given the importance of this question, it would be helpful if published papers provided summary data on a range of potentially appropriate analyses, a situation uncommon in current practice. [44]

The best time for an experimenter to think about all these issues is at the design stage of the experiment, as fundamental design flaws can never be satisfactorily corrected through analysis.

BLOCK DESIGNS

Interventions aiming to help health professionals provide better health care are subject to variation at a number of levels, both in their analysis and subsequent application. The findings of implementation studies arise from particular health-care contexts (i.e. they present particular messages in particular settings), but may not apply well in different contexts. If the message changes, will the method of behavioural change still achieve the same result? In general, there are a number of issues concerning variation and bias that are particular to health-professional behaviour-change experiments, and these naturally lead to a discussion of block designs. Block designs offer not only the potential for increased precision, but also the ability to separate one source of variation from another. For instance, a study examining the impact of interventions on different topics in the *same* health professionals may provide particularly useful information on the generalisability of an intervention across topics.

The Evidenced-Based OutReach (EBOR) trial, in progress at the time of writing, aims to evaluate the use of academic detailing in improving uptake of the findings of evidence-based cost-effectiveness guidelines. General practices in six pairs of health authorities are randomised to receive either mailed guidelines or outreach visits in four clinical topics (See Figure 3). Here, there are six ways of dividing the four topics into two 'active' and two untargeted topics (which serve as controls), and all six appear in the experiment: one in each of the six pairs of health authorities.

Figure 3: The EBOR study design

	Health Authorities					
	1 North 1 South	2 North 2 South	3 North 3 South	4 North 4 South	5 North 5 South	6 North 6 South
Ace inhibitors	Outreach	Outreach	Outreach	Mail	Mail	Mail
Antidepressants	Outreach	Mail	Mail	Outreach	Outreach	Mail
Antiplatelet therapy	Mail	Outreach	Mail	Outreach	Mail	Outreach
NSAIDs	Mail	Mail	Outreach	Mail	Outreach	Outreach

The EBOR trial is a real-world, randomised study that will provide internally and externally valid estimates of the effect of educational outreach visits by trained community pharmacists, intended to improve the quality of prescribing in English primary care. The complex design of the study is not experienced by the individual health authorities taking part, since selected practices simply receive two guidelines by outreach. The design has substantial advantages over a simple randomised design, since a range of guidelines are included and there is replication at the level of the health authority, meaning that a single authority provides only one-twelfth of the data for an individual guideline.

Such designs are sometimes regarded as balanced, incomplete block designs (BIBD), [45] with 'incompleteness' due to the inherent *impossibility* of including all interventions within a block, rather than the *undesirability* of so doing. More cogently, if the topics used are considered representative of some larger set of candidate topics to which one would wish to generalise the results, it is better to regard them as a blocking factor and the plan a row and column design, rather than a BIBD.

The Hawthorne effect, (the effect on behaviour simply as a result of involvement in a study) is a potential bias in studies that treat controls and active intervention groups differently. It has been argued that, as all subjects in a block design receive active interventions, the Hawthorne effect is likely to be equal for all those involved in such a study and the measure of the effect of the intervention will not be biased. [46] Block designs by themselves should not be expected to provide a complete answer to this problem. The assumption that a Hawthorne effect acts in a uniform manner across all activities for the health professional (for active intervention topics as well as control topics) is probably unrealistic. The beneficial effect of being part of a trial may be larger for active topics than hidden or untargeted ones and it is natural to think that this will be the case if the Hawthorne effect is regarded as a form of placebo effect that cannot be addressed by using a placebo control (as it would be in a drug trial). Further, even if the Hawthorne effect has some overall, uniform, non-specific effect upon study subjects, it can still interact with the intervention. The potential impact upon generalisability should not be ignored. [47] In other words, the study context needs to be considered as part of the intervention received. It may be that effects identified in a study would not have occurred had the Hawthorne effect not been present, or that the Hawthorne effect may ameliorate the positive impact of an intervention, and effects in practice would be larger. Put another way, a sample that may have begun as random (of GP practices, say) and so generalisable to some larger population may no longer be random as soon as it becomes the subject of experiment, if the experiment is in some way 'intrusive'. Thus, although it is possible that designs in which all subjects receive active interventions have the best chance of avoiding bias through Hawthorne effects, it also remains important that investigators try to reduce the impact of involvement in an investigation in other ways.

The problem of 'contamination' is another concern in block designs. By including both active and control interventions in the same block, there is the assumption that behavioural interventions have effects specific to their targeted activity, and so will not act upon non-targeted interventions. Often this assumption will seem quite reasonable. Taking the example of Norton & Dempsey (1985) described above, [48] it is assumed that audit and feedback targeted at the treatment of vaginitis will not impact upon the treatment of cystitis by the same health professional, and vice versa. In other cases, the possibility of contamination may be greater. Where, for instance, the investigators are examining the impact of educational outreach visits on health professionals' uptake of guidelines, it may be that some aspects of the educational process will impact upon the uptake of guidelines generally, leading to systematic underestimation of the treatment effect.

Weighing the pros and cons of block designs

Block designs offer potential advantages in the reduction of both variance and bias, as discussed above, (accepting the caveat concerning contamination). They can also be made quite complex, simultaneously providing answers to several questions of interest within the same context. This may improve the generalisability of a study and also add to its 'construct validity'.

Block designs (in which all subjects receive an active intervention) may limit the impact of the Hawthorne effect, but it is sometimes possible, both ethically and practically, to conduct trials where subjects are totally unaware that they are part of an experiment providing a complete answer to the Hawthorne effect. An example is the trial by Oakeshott et al, [49] where 170 UK family physicians in 62 practices were randomised to receive guidelines on the appropriateness of requests for radiological examinations, or control. Outcome measures were derived from routinely collected data sources and the trial addressed the important question of whether or not it is worthwhile mailing guidelines to local practitioners. Although a block design would have been possible, where different radiological guidelines were sent to different groups, this would have addressed a slightly different question (specifically, the content of the guidelines) which may not have been the object of the research. Given the availability of routinely collected outcome data, the facility to run the trial in the background and the nature of the question addressed, in this instance, the simple design used had clear advantages over more complex alternatives.

In general, simplicity in design will bring the advantage of a 'light touch' and less room for error. If the pertinent questions can be addressed within a simple design, the decision between a simple or complex design can be made merely on the grounds of which is the more cost-effective in gaining the desired information. Addressing the right questions is vital, but it will not always be the case that complexity of design helps us do this. Whilst complex designs may be useful for examining the interactions between different parts of an intervention, they may also, by virtue of their sophistication, prove difficult to interpret. [50] Additionally, the artificial constraints inherent in such designs (particularly the restriction of the Latin square that the numbers of treatments must be equal to the numbers of rows and columns of the square) may impede the optimum choice of treatments included in the study. Simple designs may be useful to examine simple questions, but they may be equally good for rather complicated ones - an example being the influence of opinion leaders on rates of caesarean section in Canada. [51]

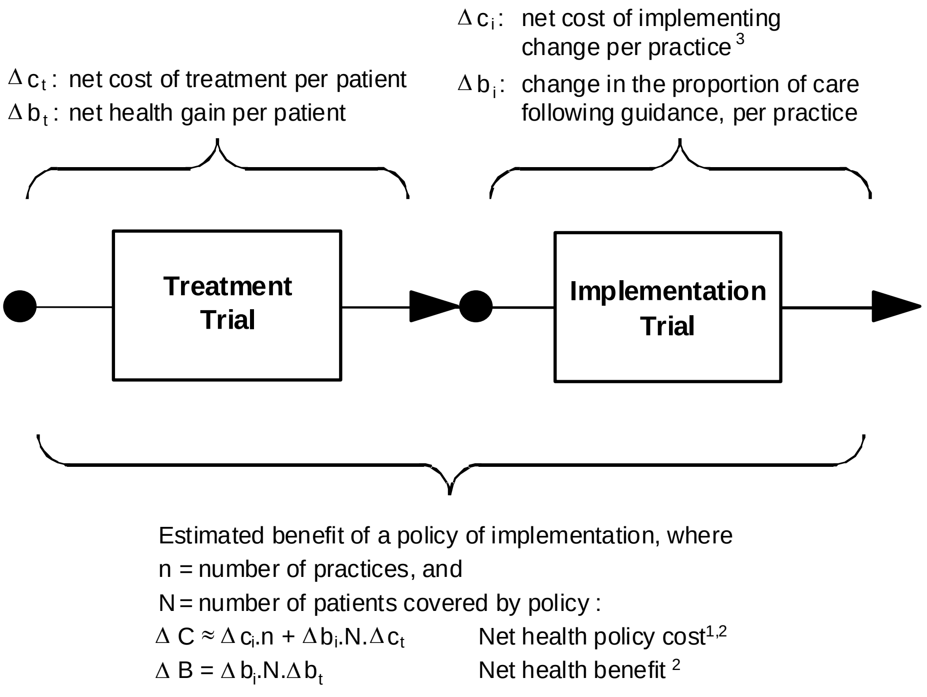
AN ECONOMIC PERSPECTIVE

In an earlier section, it was emphasised that groundwork in understanding the nature of an intervention is necessary before designing and conducting trials. This is true not only of the method used but also of the message. Representatives of all relevant parties need to agree on the necessity and achievability of a change in activity before implementation is attempted. If the need for change has not been established by available evidence, then the result of an implementation study will be uninterpretable, reflecting both varying response to the method and varying scepticism about the message.

Outcomes

It is normally appropriate to use intermediate outcomes to estimate the effect of an implementation strategy (e.g. the proportion of patients per clinical practice for whom prescribed care follows guidelines recommendations). An implementation study, targeted at changing practice in line with good evidence of effectiveness and cost-effectiveness from treatment trials, may permit health service decision-makers to evaluate a policy of changing professional behaviour by combining the implementation findings with treatment trial results or, if appropriate, a meta-analysis (Figure 4). Thus the implementation trial measures the change in resources required by the implementation method and level of change in clinician behaviour achieved (Δb_i, Δc_i). These findings are combined with the health gain estimates and resource implications from trials, which have previously assessed treatment (Δb_t, Δc_t). Evaluating a policy of behaviour change requires the decision-maker to include scale factors specific to the local context (n, the number of practices affected; N, the number of patients).

Figure 4: Evaluating the decision to influence professional behaviour.


Δc_i: net cost of implementing change per practice [3]
Δc_t: net cost of treatment per patient
Δb_t: net health gain per patient
Δb_i: change in the proportion of care following guidance, per practice
Treatment Trial
Implementation Trial
Estimated benefit of a policy of implementation, where
n = number of practices, and
N = number of patients covered by policy :
$\Delta C \approx \Delta c_i.n + \Delta b_i.N.\Delta c_t$ Net health policy cost[1,2]
$\Delta B = \Delta b_i.N.\Delta b_t$ Net health benefit [2]

1 This may serve as a useful starting point, but formally a local recosting exercise is required
2 Assuming the treatment trial has current care as a comparison to the intervention
3 Or the cost and effect of implementation may be analysed at the level of the clinician, in which case n becomes the number of clinicians

A common mistake made by health economists new to the field of implementation research is to recite the mantra that patient health outcome data should be measured: this is unlikely to be appropriate in implementation trials. Appropriate statistical analysis would be complex, since the outcomes of groups of patients would be related where they shared the same clinician or practice. Additionally, costly collection of a much larger data set would generally be required to record patient outcomes, as well as a larger trial of longer duration. For these reasons, implementation trials measuring patient health outcomes are likely to fall between stools, failing to reach statistical significance in health gains and confusing doctors as to why health outcome is being measured at all if the benefits are already proven.

Where patient outcome measures are argued for, as part of the design of an implementation trial, this issue should be explored and if there is concern that benefits from trials may not translate into local practice a pragmatic treatment trial may be required instead (i.e. the worthwhile nature of intervention has yet to be fully established). There is a more general viewpoint that health outcomes should be measured routinely in clinical practice, but implementation trials may not be a suitable vehicle to introduce clinical audit using health outcomes, since the respective data requirements are quite different.

The perspective of the economics of an implementation trial is legitimately narrow. The trial should be adequately powered to assess clinically important levels of behaviour change and be set against the net cost of implementation. The resources required to achieve change should be reported in detail so that decision-makers can cost implementation in their own locality.

Choosing alternatives

Interventions may take the form of passively disseminated printed educational materials, computer support systems, active implementation involving inclusion in focused educational events, audit cycles, discussion with a pharmacist or changes in financial incentive structures. Choosing relevant alternatives in trials to compare with implementation strategies requires care and a range of potential alternatives may be appropriate. Since it is common for there to be no formal implementation of specific research findings, it is generally appropriate to compare an implementation strategy with no formal intervention. The control group may receive no formal intervention, but be exposed to the same material by publication in professional journals (the traditional path of implementation) or may be handed the materials in a passive implementation as a means of evaluating the added benefits of a more active approach.

When considering alternatives, it may be invaluable to solicit the perspective of health service decision-makers for their views of the relevant alternatives and to address whether more active and expensive implementation strategies stand any chance of being adopted. [52] Costs and consequences may be expected to vary substantially, depending upon the clinical intervention chosen, the health system being studied, and the implementation approach used. (Printed educational materials are relatively cheap to distribute; focused educational meetings are much more expensive.) It is often difficult to transfer resources within health systems and changes in practice effected by guidance may have apparently perverse results that should be considered. For example, it has been argued that reducing the level of surgery for otitis media with effusion in children in the UK will, overall, lead to an increase rather than a decrease in costs, since available theatre time may be used for interventions that are more resource intensive and probably equally poorly supported by the evidence. [53]

Choosing a power provider

A successful implementation trial would contain both a message to convince clinicians of the need for change, and achieve a level of change that managers consider worthwhile. However, a 'clinically important change' for an implementation method is not yet a well-defined concept and requires further study. Some of the issues for clinicians and managers are rehearsed here.

It is uncertain how health policy decision-makers will respond to the results of implementation studies. Will they be interested simply in the proportion of change achieved by the trial (for example: 1% change - bad, 10% change - good) conveying a 'worthwhile' message, will they respond to the overall benefit (proportion of change multiplied by health gain), or will the overall budgetary implications of change be most influential? As a starting point for powering implementation trials, it may be appropriate to assume that decision-makers (like clinicians) respond to relatively simple notions like 'worthwhile' rather than more complex constructs like 'cost-effective' or 'efficient'. Hence, the primary objective may be set by identifying the level of change in behaviour that managers perceive to be important enough to act upon, though there are risks in this approach. It may be that relatively cheap implementation methods, achieving small changes in practice, provide the most cost-effective strategy to achieve worthwhile health gains, albeit remaining unattractive to managers. Similarly, there may be situations where less cost-effective methods may achieve greater absolute changes in the treatment of patients.

Design of economic aspects of implementation studies needs care, as does the presentation of economic messages in guidance given to clinicians. Health economists naturally equate a 'worthwhile' message with 'cost-effectiveness', but it remains a research issue as to whether clinicians respond to cost-effectiveness *per se* or to other constructs (e.g. first: this treatment has important health benefits; second: I can achieve these benefits; third: the budget can afford it). It may be important to consider the nature and presentation of the message in an implementation study with respect to the size of costs, benefits and other attributes of treatment, as all of these may influence the uptake of a message.

For example, low-dose aspirin prophylaxis for patients with cardiovascular disease is very cheap and likely to be relatively cost-effective when compared with ace-inhibitors given for heart failure, but ace-inhibitors are likely to have substantially greater health benefits. Additionally, the time taken to achieve health benefits can vary considerably. In economic evaluation, the traditional approach for coping with differential timing of costs and benefits is to discount future (quality-adjusted) survival and cost profiles. Discounting (and the whole health outcome metric) may inadequately reflect clinician preference for treatments with relatively immediate and substantial benefits measured directly from trials of treatments, as opposed to primary prevention activities with a long lead time to benefits (often) estimated (modelled) from a poorer evidence base.

DISCUSSION

The successful introduction of a new pharmaceutical product follows from a long and painstaking process with substantial attrition of 'good ideas' *en route*. Similar rigour in design of implementation studies should reduce the enactment of poorly conceived, but expensive, studies that have skipped the preliminary stages without adequate consideration. Clarity on the aims of an experiment is vital in determining a sensible protocol. Questions of different types can rarely be addressed efficiently in a single experiment and it is a mistake to try. A structured development process is required for implementation methods, assisted by a sequence of different sorts of study or experiment. The 'explanatory/ pragmatic' model of experiments is often useful as a guide in the latter stages of development, but there is no substitute for examining each new case as it arises, carefully defining the aims, and being guided by these and the practical constraints that apply.

Explanatory trials take care to control or measure extraneous factors that might influence the treatment effect, whilst in pragmatic trials, close control is neither possible nor desirable. As a principle, it remains invaluable to establish that an intervention can work and to gain an understanding of the way it works in studies with high construct validity before attempting to evaluate whether it can achieve this potential in the real world.

Although useful, non-equivalent group designs require more careful interpretation than randomised studies. A number of potential biases have been discussed: scaling differences, maturational processes, cyclical effects, selection and regression to the mean.

An inappropriate choice of unit of analysis in implementation studies is at best inefficient, and at worst deceptive. Analysing at the level of patient rather than physician is inappropriate wherever we are interested in physician behaviour that cannot be explained by patient variability; in other words, where there are likely to be important differences between physicians. This is not an empirical issue. Fitting a model that ignores physician variability in these circumstances is akin to expecting a free lunch (and there is no such thing). The appropriate unit of analysis for implementation trials is the changed behaviour achieved in clinicians, either individually or as a practice group.

Whilst more sophisticated statistical modelling embracing several sources of variation is possible, a relatively straightforward analysis at the right level may well retrieve most of the information from the data. Multi-level models using appropriate statistical techniques can have some advantages over simple analyses at the appropriate unit of analysis, especially where antecedence data are available that may explain some of the variability observed. However, more work is required, particularly in the context of controlled trials.

The use of a block design in which individual practices provide data both on the effect of outreach and control reduces variability in outcome measures, providing more precise estimates of treatment effect. The approach also extends the scope of the evaluation from a single guideline topic to a range of guidelines, thus increasing the generalisability of the study.

Implementation methods should not be considered before it has been established that a worthwhile treatment or health care intervention is being under-utilised. Implementation trials conducted without this foundation may fail either because the method is ineffective or because clinicians do not accept the importance of the message. Careful consideration is required in the design of these trials to ensure evaluation of the implementation method and avoid confounding influences.

The resource consequences of the implementation method and the level of behavioural change achieved are being measured in an implementation trial, not the costs of treatment or health outcomes. Existing evidence of the cost-effectiveness of treatment should have contributed to the basis for the need of an implementation strategy. With appropriate methodology, it appears possible to design studies which ‘bolt on’ to the treatment evidence and allow decision-makers to assess the value of a policy to implement change in their own setting. However, whether those involved in policy and treatment decisions respond to levels of change in behaviour or more complex constructs of health benefits or cost-effectiveness resulting from change remains a research issue.

Educational outreach visits by trained pharmacists are a commonly used strategy intended to affect prescribing practice in the UK, though to date there have been no large randomised trials examining the efficiency of such practices. EBOR will provide a valid estimate of the effect of outreach based upon a random sample of practices.

REFERENCES

1. Secretary of State for Health. The new NHS: modern, dependable. Cm 3807, December 1997.
2. The SOLVD investigators. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. N Engl J Med 1991;325:293-302
3. Garg R, Yusuf S. Overview of randomized trials of angiotensin-converting enzyme inhibitors on mortality and morbidity in patients with heart failure. Collaborative Group on ACE Inhibitors Trials. JAMA 1995;273:1450-6.
4. Antiplatelet Trialists' Collaborative Group. Collaborative overview of randomised controlled trials of antiplatelet therapy - I: Prevention of death, myocardial infarction and stroke by prolonged antiplatelet therapy in various categories of patients. BMJ 1994;308:81-106
5. Juul-Möller S, Edvardsson N, Jahnmatz B, Rosén A, Sørensen S, Ömblus R for the Swedish Angina Pectoris Aspirin Trial (SAPAT) Group. Double-blind trial of aspirin in primary prevention of myocardial infarction in patients with stable chronic angina pectoris. Lancet 1992;340:1421-5.
6. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. Randomised trial of intravenous streptokinase, oral aspirin, both or nether among 17187 cases of suspected acute myocardial infarction: ISIS-2. Lancet 1988;ii:349-60.
7. Fibrinolytic Therapy Trialists' Collaborative Group. Indications for fibrinolytic therapy in suspected acute myocardial infarction: Collaborative overview of early mortality and major morbidity results from all randomised trials of more than 1000 patients. Lancet 1994;343:311-22.
8. Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. JAMA 1992;268:240-8.
9. Effective Health Care. Implementing Clinical Practice Guidelines. Bulletin No 8. Leeds: University of Leeds, 1994.
10. Freemantle N, Harvey E, Grimshaw J, Wolf F, Oxman A, Grilli R et al. The effectiveness of printed educational materials in changing the behaviour of healthcare professionals. In: Freemantle N, Bero L, Grilli R, Grimshaw J, Oxman A (eds.) Effective Professional Practice Module. In: The Cochrane Database of Systematic Reviews. In press
11. Rogers EM. Diffusion of innovations. New York: Free Press, 1983.
12. Eddy DM. A manual for assessing health practices and designing practice policies: the explicit approach. Philadelphia: American College of Physicians, 1992.
12. Peckham M. Research and development for the National Health Service. Lancet 1991;338:367-71
13. Department of Health. Methods to promote the implementation of research findings in the NHS - priorities for evaluation. Report to the NHS Central Research and Development Committee. Leeds, Department of Health, 1995.
15. Evans CE, Haynes RB, Gilbert JR, Taylor DW, Sackett DL, Johnston M. Educational package on hypertension for primary care physicians. Can Med Assoc J. 1984;130:719-22.
16. Evans CE, Haynes RB, Birkett NJ, Gilbert JR, Taylor DW, Sackett DL, et al. Does a mailed continuing program improve physician performance? results of a randomized trial in anti-hypertensive care. JAMA 1986;255:501-4.
17. Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? Annals of Internal Medicine 1985;103:596-9.
18. Pocock SJ. Clinical trials, a practical approach. Chichester: Wiley, 1983.
19. Soumerai SB, Avorn J. Principles of educational outreach ('academic detailing') to improve clinical decision making. JAMA 1990;263:549-556.
20. Cochrane Collaboration on Effective Professional Practice (CCEPP) Register of Studies. York, University of York, 1996.
21. Cook TD, Campbell DT. Quasi-experimentation: design and analysis issues for field settings. Chicago: Rand McNally, 1979
22. Wagner EH, Barrett P, Barry MJ, Barlow W, Fowler FJ. The Effect of a Shared Decision-Making Program on Rates of Surgery for Benign Prostatic Hyperplasia. Medical Care 1995;33 (8):765-70
23. Angell M. Patients' preferences in randomized clinical trials. N Engl J Med 1984;310:1385-7
24. Domenighetti G, Luraschi P, Casabianca A, Gutzwiller F, Spinelli A, Pedrinis E et al Effect of information campaign by the mass media on hysterectomy rates. Lancet 1988;ii:1470-3
25. Wallack L, Dorfman L, Jernigan D, Themba M. Media advocacy and public health. Power for prevention. London: Sage, 1993

26. Rice N, Leyland A. Multilevel models: applications to health data. Journal of Health Services Research and Policy 1996; 1: 154-64
27. Rice N, Jones A. Multilevel models and health economics. Health Economics 1997; 6: 561-75.
28. Shipley MJ, Smith PG, Dramaix M. Calculation of power for matched pair studies when randomization is by group. International Journal of Epidemiology 1989;18:457-61
29. Franks P, Dickson JC. Comparison of family physicians and internists: process and outcome in adult patients at a community hospital. Medical Care 1986; 24: 941-8.
30. Cornfield J. Randomization by group: a formal analysis. American Journal of Epidemiology 1978:108:100-2
31. Donner A. A Regression Approach to the Analysis of Data Arising from Cluster Randomisation. International Journal of Epidemiology 1985;14:322-326
32. Donner A, Birkett N, Buck C. Randomization by cluster: sample size requirements and analysis. American Journal of Epidemiology 1981;114:906-14.
33. Donner A, Donald A. Analysis of data arising from a stratified design with the cluster as unit of randomization. Statistics in Medicine 1987;6:43-52.
34. Donner A. Statistical methodology for paired cluster designs. American Journal of Epidemiology 1987;126:972-9.
35. Hsieh FY. Sample size formulae for intervention studies with the cluster as unit of randomization. Statistics in Medicine 1988;8:1195-201.
36. Goldstein, H, Multilevel models in educational and social research. 1987. Oxford University Press, London.
37. Duffy SW, South MC, Day NE. Cluster randomisation in large public health trials: the importance of antecedence data. Statistics in Medicine 1992; 307-16.
38. Langford IH, Bentham G. Regional variations in mortality rates in England and Wales: an analysis using multi-level modeling. Social Science in Medicine 1996; 42: 897-908.
39. Feng Z, McLerran D, Grizzle J. A comparison of statistical methods for clustered data analaysis with Gaussian error. Statistics in Medicine 1996; 15: 1973-806.
40. Diwan VK, Eriksson B, Sterky G, Tomson G. Randomization by group in studying the effect of drug information in primary care. Int J Epidemiol 1992 Feb;21:124-130.
41. Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random-effects meta analysis: a comparative study. Stats in Med 1995; 14: 2685-99.
42. Davis P, Gribben B. Rational prescribing and interpractitioner variation: a multilevel approach. International Journal of Technology Assessment in Health Care 1995;11:428-42.
43. Van de Werf F, Topol EJ, Lee KL, Woodlief LH, Granger CB, Armstrong PW et al. Variations in patient management and outcomes for acute myocardial infarction in the United States and other countries: Results from the GUSTO trial. JAMA 1995;273:1586-91.
44. Freemantle N, Haines A, Mason JM, Eccles M. CONSORT - an important step towards evidence based health care. Annals of Internal Medicine, 1997;126:81-3
45. Cochran WG, Cox GM. Experimental designs. Chirchester: Wiley, 1992
46. Russell I, Grimshaw J. The effectiveness of referral guidelines: a review of the methods and findings of published evaluations. In: Roland M, Coulter A eds. Hospital Referrals. Vol 1, Oxford: Oxford University Press, 1992; 179-211
47. Freemantle N, Drummond MF. Should clinical trials with concurrent economic analyses be blinded? JAMA 1997;277: 63-4
48. Norton PG, Dempsey LJ. Self-audit: Its effect on quality of care. J Fam Pract 1985;21: 289-91
49. Oakeshott P, Kerry SM, Williams JE. Randomized controlled trial of the effect of the Royal College of Radiologists' guidelines on general practitioners' referrals for radiographic examination. Br J Gen Pract 1994;44: 197-200
50. North of England Study of Standards and Performance in General Practice. Medical audit in general practice: effects on doctors' clinical behaviour and the health of patients with common childhood conditions. BMJ 1992;304: 1480-8
51. Lomas J, Enkin M, Anderson GM, Hannah WJ, Vayda E, Singer J. Opinion leaders vs audit and feedback to implement practice guidelines. Delivery after previous cesarean section. JAMA 1991; 265: 2202-7
52. Freemantle , Wood J, Crawford F. Evidence into practice, experimentation and quasi-experimentation: are the methods up to the task? Journal of Epidemiology and Community Health 1998; 52: 75-81.
53. Freemantle N, Watt I, Mason J. Developments in the purchasing process in the NHS: Towards an explicit politics of rationing. Public Administration 1993; 71: 535-48.