MRC Medical Research Council

Developing and evaluating complex interventions:

new guidance

Prepared on behalf of the Medical Research Council by:

Peter Craig, MRC Population Health Sciences Research Network
Paul Dieppe, Nuffield Department of Orthopaedic Surgery, University of Oxford
Sally Macintyre, MRC Social and Public Health Sciences Unit
Susan Michie, Centre for Outcomes Research and Effectiveness, University College London
Irwin Nazareth, MRC General Practice Research Framework
Mark Petticrew, Department of Public Health and Policy, London School of Hygiene and Tropical Medicine

www.mrc.ac.uk/complexinterventionsguidance

Contents

Acknowledgements

Many people have contributed to the preparation of this document. The writing group would like to thank participants in the May 2006 workshop for initiating the revision of the 2000 Guidelines, and the MRC Health Services and Public Health Research Board which provided financial support.

We should also like to thank the following for their helpful comments on drafts: Professor Sheila Bird, MRC Biostatistics Unit; Professor Iain Crombie, University of Dundee, Professor Janet Darbyshire, MRC Clinical Trials Unit, Professor Peter Diggle, University of Lancaster, Professor Martin Eccles, University of Newcastle, Dr Liz Fenwick, University of Glasgow, Professor Ian Harvey, University of East Anglia, Dr Jill Francis, University of Aberdeen, Professor Penny Hawe, University of Calgary, Professor Hazel Inskip, MRC Epidemiology Resource Centre, Professor Marie Johnston, University of Aberdeen, Professor Mike Kelly, National Institute for Health and Clinical Excellence, Professor Paul Little, University of Southampton, Professor Laurence Moore, University of Cardiff, Professor Max Parmar, MRC Clinical Trials Unit, Professor Jennie Popay, University of Lancaster, Dr Sasha Shepperd, University of Oxford, Professor Alan Shiell, University of Calgary, Dr Danny Wight, MRC Social and Public Health Sciences Unit, Professor Lucy Yardley, University of Southampton, and Dr Pat Yudkin, University of Oxford. In doing so we do not wish to imply that they endorse this document.

Summary

This document provides guidance on the development, evaluation and implementation of complex interventions to improve health. It updates the advice provided in the 2000 *MRC Framework for the Development and Evaluation of RCTs for Complex Interventions to Improve Health,* taking account of the valuable experience that has accumulated since then, and extending the coverage in the guidance of non-experimental methods, and of complex interventions outside the health service. It is intended to help researchers to choose appropriate methods, research funders to understand the constraints on evaluation design, and users of evaluation to weigh up the available evidence in the light of these methodological and practical constraints. Box 1 summarises the main elements of the process, and the key questions that researchers should ask themselves as they work through it.

Box 1 The development-evaluation-implementation process

Developing, piloting, evaluating, reporting and implementing a complex intervention can be a lengthy process. All of the stages are important, and too strong a focus on the main evaluation, to the neglect of adequate development and piloting work, or proper consideration of the practical issues of implementation, will result in weaker interventions, that are harder to evaluate, less likely to be implemented and less likely to be worth implementing.

Developing an intervention

Questions to ask yourself include: Are you clear about what you are trying to do: what outcome you are aiming for, and how you will bring about change? Does your intervention have a coherent theoretical basis? Have you used this theory systematically to develop the intervention? Can you describe the intervention fully, so that it can be implemented properly for the purposes of your evaluation, and replicated by others? Does the existing evidence – ideally collated in a systematic review – suggest that it is likely to be effective or cost effective? Can it be implemented in a research setting, and is it likely to be widely implementable if the results are favourable?

If you are unclear about the answers to these questions, further development work is needed before you begin your evaluation. If you are evaluating a policy or a service change as it is being implemented, rather than carrying out an experimental intervention study, you still need to be clear about the rationale for the change and the likely size and type of effects, in order to design the evaluation appropriately.

Piloting and feasibility

Questions to ask yourself include: Have you done enough piloting and feasibility work to be confident that the intervention can be delivered as intended? Can you make safe assumptions about effect sizes and variability, and rates of recruitment and retention in the main evaluation study?

Evaluating the intervention

Questions to ask yourself include: What design are you going to use, and why? Is an experimental design preferable and if so, is it feasible? If a conventional parallel group randomised controlled trial is not possible, have you considered alternatives such as cluster randomization or a stepped wedge design? If the effects of the intervention are expected to be large or too rapid to be confused with secular trends, and selection biases are likely to be weak or absent, then an observational design may be appropriate. Have you set up procedures for monitoring delivery of the intervention, and overseeing the conduct of the evaluation?

Including a process evaluation is a good investment, to explain discrepancies between expected and observed outcomes, to understand how context influences outcomes, and to provide insights to aid implementation. Including an economic evaluation will likewise make the results of the evaluation much more useful for decision-makers.

Reporting

Questions to ask yourself include: Have you reported your evaluation appropriately, and have you updated your systematic review? It is important to provide a detailed account of the intervention, as well as a standard report of the evaluation methods and findings, to enable replication studies, or wider scale implementation. The results should ideally be presented in the context of an updated systematic review of similar interventions.

Implementation

Questions to ask yourself include: Are your results accessible to decision-makers, and have you presented them in a persuasive way? Are your recommendations detailed and explicit?

Strategies to encourage implementation of evaluation findings should be based on a scientific understanding of the behaviours that need to change, the relevant decision-making processes, and the barriers and facilitators of change. If the intervention is translated into routine practice, monitoring should be undertaken to detect adverse events or long term outcomes that could not be observed directly in the original evaluation, or to assess whether the effects observed in the study are replicated in routine practice.

Introduction

Complex interventions are widely used in the health service, in public health practice, and in areas of social policy such as education, transport and housing that have important health consequences. Conventionally defined as interventions with several interacting components, they present a number of special problems for evaluators, in addition to the practical and methodological difficulties that any successful evaluation must overcome. Many of the extra problems relate to the difficulty of standardising the design and delivery of the interventions,[1 2] their sensitivity to features of the local context,[3 4] the organisational and logistical difficulty of applying experimental methods to service or policy change,[5 6] and the length and complexity of the causal chains linking intervention with outcome.[7]

In 2000, the MRC published a *Framework for the Development and Evaluation of RCTs for Complex Interventions to Improve Health*[8], to help researchers and research funders to recognise and adopt appropriate methods. The guidelines have been highly influential, and the accompanying BMJ paper[9] is widely cited. In May 2006 a workshop was held under the auspices of the MRC Population Health Sciences Network to review the 2000 Framework[10]. It was attended by several members of the group that developed the Framework, and others with an interest in the evaluation of complex interventions. Participants identified a number of limitations in the 2000 Framework, including: (1) the adoption of a model based on the phases conventionally used in the evaluation of new drugs, and the linearity implied by the diagram of the model; (2) the lack of evidence for many of the recommendations; (3) the limited guidance on how to approach developmental and implementation phase studies; (4) an assumption that conventional clinical trials provide a template for all the different approaches to evaluation; (5) a lack of guidance on how to tackle highly complex or non-health sector interventions, e.g. programmes made up of several complex interventions; and (6) the lack of attention to the social, political or geographical context in which interventions take place. The workshop concluded that, while the Framework remained valuable, much useful experience had accumulated since it was published that should be incorporated in a revised version.

The following guidance is designed to update and extend the 2000 Framework. It re-emphasises some of the key messages, but also tries to address the limitations identified by (1) providing a more flexible, less linear model of the process, giving due weight to the development and implementation phases, as well as to evaluation, and (2) giving examples of successful approaches to the development and evaluation of a variety of complex interventions, using a range of methods from clinical trials to natural experiments. While some aspects of good practice are clear, methods for developing, evaluating and implementing complex interventions are still being developed, and on many important issues there is no consensus yet on what is best practice.

The guidance is presented as follows:

In part I, we set out the revised framework and the key messages to emerge from recent work on complex interventions.

In part II, we suggest some further questions that researchers considering a complex intervention study should ask themselves.

In part III we present a range of case studies, using a variety of study designs, carried out in a wide range of settings.

Some of the issues we discuss are specific to complex interventions; others are more general, but are included because they apply with particular force to the development and evaluation of complex interventions.

The guidance is primarily intended to help researchers choose and implement appropriate methods, given the state of existing knowledge and the nature of their target intervention. We hope that it will also help research funders to understand the constraints on evaluation design and recognise appropriate methodological choices. Finally, we hope that it will enable policy makers, practitioners and other commissioners and users of evaluation to weigh up the available evidence in the light of these methodological and practical constraints, and to consider carefully how their own decisions affect the quality of the evidence that evaluation of complex interventions can provide.

Part 1 Key messages

(1) What makes an intervention complex?

Complex interventions are usually described as interventions that contain several interacting components. There are, however, several dimensions of complexity: it may be to do with the range of possible outcomes, or their variability in the target population, rather than with the number of elements in the intervention package itself. It follows that there is no sharp boundary between simple and complex interventions. Few interventions are truly simple, but there is a wide range of complexity. Box 2 summarises some of the dimensions of complexity and their implications for developing and evaluating interventions.

Box 2 What makes an intervention complex?

Some dimensions of complexity

- Number of and interactions between components within the experimental and control interventions
- Number and difficulty of behaviours required by those delivering or receiving the intervention
- Number of groups or organisational levels targeted by the intervention
- Number and variability of outcomes
- Degree of flexibility or tailoring of the intervention permitted

Implications for development and evaluation

- A good theoretical understanding is needed of how the intervention causes change, so that weak links in the causal chain can be identified and strengthened
- Lack of impact may reflect implementation failure (or teething problems) rather than genuine ineffectiveness; a thorough process evaluation is needed to identify implementation problems.
- Variability in individual level outcomes may reflect higher level processes; sample sizes may need to be larger to take account of the extra variability, and cluster- rather than individually-randomized designs considered.
- Identifying a single primary outcome may not make best use of the data; a range of measures will be needed, and unintended consequences picked up where possible.
- Ensuring strict fidelity to a protocol may be inappropriate; the intervention may work better if adaptation to local setting is allowed.

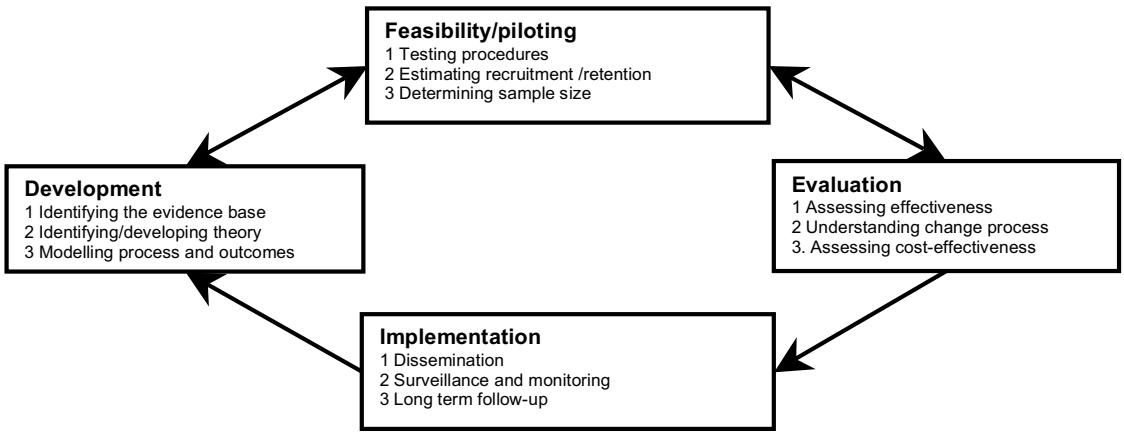
How you deal with complexity will depend on the aims of the evaluation. A key question in evaluating a complex intervention is about practical effectiveness – whether the intervention works in everyday practice[11] – in which case it is important to understand the whole range of effects, how they vary among recipients of the intervention, between sites, over time, etc., and the causes of that variation. In the case of trials, the appropriate comparison will usually be with 'usual treatment' rather than a placebo. Once effectiveness has been established, the focus of interest may shift towards fine-tuning the intervention, or delivering it more efficiently, in which case it may be possible to draw a more straightforward comparison by allowing one key element of the intervention to vary, while balancing the other elements between arms of a trial.

A second key question in evaluating complex interventions is *how* the intervention works, in other words, what are the active ingredients within the intervention and how are they exerting their effect?[12] Only by addressing this kind of question can we build a cumulative understanding of causal mechanisms, design more effective interventions and apply them appropriately across group and setting.

(2) The development-evaluation-implementation process

The process from development through to implementation of a complex intervention may take a wide range of different forms. Figure 1 summarises the main stages and the key functions and activities at each stage. The arrows indicate the main interactions between the phases. Although it is useful to think in terms of stages, often these will not follow a linear or even a cyclical sequence.[13] Reporting is not shown as a separate activity, because we suggest it should be seen as an important element of each stage in the process (see part II, Questions 3 and 9).

Figure 1 Key elements of the development and evaluation process



Best practice is to develop interventions systematically, using the best available evidence and appropriate theory, then to test them using a carefully phased approach, starting with a series of pilot studies targeted at each of the key uncertainties in the design, and moving on to an exploratory and then a definitive evaluation. The results should be disseminated as widely and persuasively as possible, with further research to assist and monitor the process of implementation.

In practice, evaluation takes place in a wide range of settings that constrain researchers' choice of interventions to evaluate and their choice of evaluation methods. Ideas for complex interventions emerge from various sources, including: past practice, existing evidence, theory, an investigator, policy makers or practitioners, new technology, or commercial interests. The source may have a significant impact on how much leeway the investigator has to modify the intervention, to influence the way it is implemented, or to choose an ideal evaluation design. Sometimes, evaluation may take place alongside large scale implementation, rather than beforehand.[14 15 16] Strong evidence may be ignored or weak evidence rapidly taken up, depending on its political acceptability or fit with other ideas about what works.

Researchers need to consider carefully the trade-off between the importance of the intervention, and the value of the evidence of effectiveness that can be gathered given these constraints. For example, in evaluating the health impact of a social intervention, such as a programme of housing or neighbourhood regeneration, researchers may have no say over what the intervention consists of, and little influence over how or when the programme is implemented, limiting the scope to undertake development work or to influence allocation. Although experimental and quasi-experimental methods are becoming more widely used and accepted,[17 18] there may be political or ethical objections to randomisation in settings in which such methods are unfamiliar. Where there are significant non-health benefits associated with receipt of an intervention, the ethics of withholding or delaying receipt of an intervention in order to study its health impact need careful consideration and sensitive handling.[19 20] Given the cost of such interventions, evaluation should still be considered: 'best available' methods, even if they are not theoretically optimum, may yield useful results.[5 6]

If such methods are used, researchers should be aware of their limitations, and interpret and present the findings with due caution. Wherever possible, evidence should be combined from a variety of sources that do not all share the same weaknesses.[21] Researchers should also be prepared to explain to decision-makers the pros and cons of experimental and non-experimental approaches (see Part I, Section 5), and the trade-offs involved in settling for weaker methods. They should be prepared to challenge decision-makers when, for example, interventions of uncertain effectiveness are being implemented in a way that would make strengthening the evidence through a rigorous evaluation difficult, or when a modification of the implementation strategy would open up the possibility of a much more informative evaluation.[14 21]

(3) Developing a complex intervention

Before undertaking a substantial evaluation you should first develop the intervention to the point where it can reasonably be expected to have a worthwhile effect.

Identifying the evidence base: You should begin by identifying the relevant, existing evidence base, ideally by carrying out a systematic review (see also Part II, Question 1). You may be lucky and find a recent high quality review that is relevant to your intervention, but it is more likely that you will have to conduct one yourself, and maintain and update it as the evaluation proceeds.

Identifying/developing appropriate theory: You also need to be aware of the relevant theory, as this is more likely to result in an effective intervention than is a purely empirical or pragmatic approach.[22] The rationale for a complex intervention, i.e. what changes are expected, and how change is to be achieved, may not be clear at the outset. If so, a vitally important early task is to develop a theoretical understanding of the likely process of change, by drawing on existing evidence and theory, supplemented if necessary by new primary research, for example interviews with 'stakeholders', i.e those targeted by the intervention, or involved in its development or delivery. This should be done whether you are developing the intervention you are planning to evaluate, or evaluating an intervention that has already been developed and/or implemented. There may be lots of competing or partly overlapping theories,[23] and finding the most appropriate ones will require expertise in the relevant disciplines.[24]

Modelling process and outcomes: Modelling a complex intervention prior to a full scale evaluation can provide important information about the design of both the intervention and the evaluation (Case study 1).[25-28] One useful approach to modelling is to undertake a pre-trial economic evaluation.[29-31] This may identify weaknesses and lead to refinements, or it may show that a full-scale evaluation is unwarranted, for example because the effects are so small that a trial would have to infeasibly large (Case study 2).[32] Formal frameworks for developing and testing complex interventions, such as MOST[33] or RE_AIM,[34] may be a good source of ideas, and the National Institute for Health and Clinical Excellence has produced detailed guidance on the development and evaluation of behaviour change interventions.[35]

It is important to begin thinking about implementation at an early stage in developing an intervention and to ask the question 'would it be possible to use this?' before embarking on a lengthy and expensive process of evaluation.[36 37] You also need to ask 'by whom (national or local policy-makers, opinion leaders/formers, practitioners, patients, the public, etc.)?' and in what population or setting. Work out who needs to know about the outcome of the evaluation, and what kind of information they will require in order to implement the changes that might be indicated by the new evidence. Who (or what) are the facilitators? What (or who) are the obstacles? Why is your evidence likely to be persuasive? It may not be convincing if it conflicts with deeply entrenched values.

(4) Assessing feasibility and piloting methods

The feasibility and piloting stage includes testing procedures for their acceptability, estimating the likely rates of recruitment and retention of subjects, and the calculation of appropriate sample sizes. Methodological research suggests that this vital preparatory work is often skimped.[38] Evaluations are often undermined by problems of acceptability,[39 40] compliance,[41] delivery of the intervention, recruitment and retention,[42-44] smaller-than-expected effect sizes, and so on, that could be anticipated by thorough piloting. A pilot study need not be a 'scale model' of the planned mainstage evaluation, but should address the main uncertainties that have been identified in the development work (Case study 3). Pilot study results should be interpreted cautiously when making assumptions about the required sample size, likely response rates, etc., when the evaluation is scaled up. Effects may be smaller or more variable and response rates lower when the intervention is rolled out across a wider range of settings. A mixture of qualitative and quantitative methods is likely to be needed, for example to understand barriers to participation and to estimate response rates. Depending on the results, a series of studies may be required to progressively refine the design, before embarking on a full-scale evaluation.

(5) Evaluating a complex intervention

There are many study designs to choose from, and different designs suit different questions and different circumstances.[45] Awareness of the whole range of experimental and non-experimental approaches should lead to more appropriate methodological choices. Beware of 'blanket' statements about what designs are suitable for what kind of intervention (e.g. 'randomised trials are inappropriate for community-based interventions, psychiatry, surgery, etc.').[46] A design may rarely be used in a particular field, but that does not mean it cannot be used, and you should make your choice on the basis of specific characteristics of your study, such as expected effect size and likelihood of selection and other biases.

Assessing effectiveness: You should always consider randomisation, because it is the most robust method of preventing the selection bias that occurs whenever those who receive the intervention differ systematically from those who do not, in ways likely to affect outcomes.[47 48] If a conventional individually-randomised parallel group design is not appropriate, there are a number of other experimental designs that should be considered (Box 3).

Box 3 Experimental designs for evaluating complex interventions

Individually randomised trials: Individuals are randomly allocated to receive either an experimental intervention, or an alternative such as standard treatment, a placebo or remaining on a waiting list. Such trials are sometimes dismissed as inapplicable to complex interventions, but there are many variants of the basic method, and often solutions can be found to the technical and ethical problems associated with randomization (Case study 4).

Cluster randomised trials: Contamination of the control group, leading to biased estimates of effect size, is often cited as a drawback of randomised trials of population level interventions,[49] but cluster randomisation, widely used in health services research, is one solution. Here, groups such as patients in a GP practice or tenants in a housing scheme are randomly allocated to the experimental or a control intervention (Case study 5).

Stepped wedge designs: The randomised stepped wedge design may be used to overcome practical or ethical objections to experimentally evaluating an intervention for which there is some evidence of effectiveness, or which cannot be made available to the whole population at once. It allows a randomised controlled trial to be conducted without delaying roll-out of the intervention. Eventually, the whole population receives the intervention, but with randomisation built into the phasing of implementation (Case study 6).

Preference trials and randomised consent designs: Practical or ethical obstacles to randomisation can sometimes be overcome by the use of non-standard designs. Where patients have very strong preferences among treatments, basing treatment allocation on patients' preferences, or randomising patients before seeking consent, may be appropriate (Case study 7).

N-of-1 designs: Conventional trials aim to estimate the average effect of an intervention in a population, and provide little information about within or between person variability in response to interventions, or about the mechanisms by which effective interventions achieve change. N-of-1 trials, in which individuals undergo interventions with the order or scheduling decided at random, can be used to assess between and within person change, and to investigate theoretically predicted mediators of that change (Case study 8).

If an experimental approach is not feasible, for example because the intervention is irreversible, necessarily applies to the whole population, or because large-scale implementation is already under way, you should consider whether there is a good non-experimental alternative (Box 4). In some circumstances, randomisation may be unnecessary and other designs preferable,[50 51] but the conditions under which non-randomised designs can yield reliable estimates of effect are very limited. They are most useful where the effects of the intervention are large or rapidly follow exposure, and where the effects of selection, allocation and other biases are relatively small. Although there is a range of approaches for dealing with such biases, including conventional covariate adjustment using a regression model, and extensions such as instrumental variable[52] and propensity score[53] methods, the interpretation of small effects from non-randomised studies requires particular care and should draw on supporting evidence where possible. Such evidence might include a consistent pattern of effects across studies, such as a dose-response relationship in which more intensive variants of the interventions are associated with larger effects, or evidence from other types of study for a causal mechanism that can explain the observed effect.[21]

Box 4 Choosing between randomised and non-randomised designs

Size and timing of effects: randomisation may be unnecessary if the effects of the intervention are so large or immediate that confounding or underlying trends are unlikely to explain differences in outcomes before and after exposure. It may be inappropriate if the changes are very small, or take a very long time to appear. In these circumstances a non-randomised design may be the only feasible option, in which case firm conclusions about the impact of the intervention may be unattainable.

Likelihood of selection bias: randomisation is needed if exposure to the intervention is likely to be associated with other factors that influence outcomes. Post-hoc adjustment is a second-best solution, because it can only deal with known and measured confounders and its efficiency is limited by errors in the measurement of the confounding variables.[54 55]

Feasibility and acceptability of experimentation: randomisation may be impractical if the intervention is already in widespread use, or if key decisions about how it will be implemented have already been taken, as is often the case with policy changes and interventions whose impact on health is secondary to their main purpose.

Cost: if an experimental study is feasible, and would provide more reliable information than an observational study, you need then to consider whether the additional cost would be justified by having better information (see case study 11).

Examples where non-randomised designs have been used successfully are the evaluation of legislation to restrict access to means of suicide,[56] to prevent air pollution (Case study 9),[57 58] or to ban smoking in public places.[59-61] Non-randomised designs may have to be used for studying rare adverse events, which a trial would have to be implausibly large to detect.[62 63] An example is the use of case-control methods to evaluate the impact on the incidence of sudden infant deaths of advice about sleep position and other aspects of babies' sleeping environment.[64-66]

A crucial aspect of the design of an evaluation is the choice of outcome measures. You need to think about which outcomes are most important, and which are secondary, and how you will deal with multiple outcomes in the analysis. A single primary outcome, and a small number of secondary outcomes, is the most straightforward from the point of view of statistical analysis. However, this may not represent the best use of the data, and may not provide an adequate assessment of the success or otherwise of an intervention which may have effects across a range of domains. A good theoretical understanding of the intervention, derived from careful development work, is key to choosing suitable outcome measures. You need measures that are appropriate to the design of evaluation – subjective or self-report outcomes may be unreliable, especially if the study is unblinded. You should also consider the timing of change, and determine length of follow-up according to a clear understanding (or at least clear assumptions) of the rate and pattern of change. You may have to use surrogate outcomes, or your outcomes may act as predictors or mediators of other effects. No matter how thorough your development work, you should remain alert to the possibility of unintended and possibly adverse consequences. Finally, you need to consider which sources of variation in outcomes are important, and carry out appropriate subgroup analyses. In the case of public health interventions which are expected to impact on inequalities in health, analyses by socio-economic position may be needed.

Understanding processes: A process evaluation is often highly valuable – providing insight into why an intervention fails unexpectedly or has unanticipated consequences (Case study 10 and 13),[67] or why a successful intervention works and how it can be optimised (Case study 14). Process evaluation nested within a trial can also be used to assess fidelity and quality of implementation, clarify causal mechanisms and identify contextual factors associated with variation in outcomes.[68 69] Process evaluations should be conducted to the same high methodological standards and reported just as thoroughly as evaluation of outcomes.[70] However, they are not a substitute for an outcome evaluation, and interpreting the results is crucially dependent on knowledge of outcomes.

Assessing cost-effectiveness: An economic evaluation should be included if at all possible, as this will make the results far more useful for decision-makers. Ideally, economic considerations should be taken fully into account in the design of the evaluation, to ensure that the cost of the study is justified by the potential benefit of the evidence it will generate, appropriate outcomes are measured, and the study has enough power to detect economically important differences (Case study 11).[30 31] The main purpose of an economic evaluation is estimation rather than hypothesis testing so it may still be worth including one, even if the study cannot provide clear cost or effect differences, so long as the uncertainty is handled appropriately.[71 72]

(6) Implementation and beyond

Publication in the research literature is essential (See Part II, Questions 3 and 9), but it is only part of an effective implementation strategy. Earlier (Part I, Section 3), we stressed the need to ask relevant questions in evaluation research; here we consider what can be done to encourage uptake of results.

Getting evidence into practice: To have a chance of getting your findings translated into routine practice or policy, you need to make them available using methods that are accessible and convincing to decision-makers. It has long been recognised that passive strategies are ineffective at getting evidence into practice.[73] Information needs to be provided in accessible formats and disseminated actively.[74] The evidence base for effective implementation remains limited,[75] but some promising approaches have been identified (Box 5).

Box 5 Potentially useful approaches to implementation

Involve stakeholders in the choice of question and design of the research to ensure relevance[36 37 76]

Provide evidence in an integrated and graded way: reviews not individual studies, and variable length summaries that allow for rapid scanning[77]

Take account of context, and identify the elements relevant to decision-making, such as benefits, harms and costs[77]

Make recommendations as specific as possible[78]

Use a multifaceted approach involving a mixture of interactive rather than didactic educational meetings audit, feedback, reminders, and local consensus processes[79]

Successful implementation depends on changing behaviour – often of a wide range of people. This requires a scientific understanding of the behaviours that need to change, the factors maintaining current behaviour and barriers and facilitators to change,[24 27] and the expertise to develop strategies to achieve change based on this understanding. Further research may be needed to assist the process of implementation (Case study 12), and implementation research teams should include a behavioural scientist.

Surveillance, monitoring and long term outcomes: An experimental study is unlikely to provide a comprehensive, fully generalisable account of the effectiveness of an intervention. Few trials are powered to detect rare adverse events,[47 62] and even pragmatic studies with wide inclusion criteria are likely to take place in a population and range of settings that are to some extent self-selected. Effects are likely to be smaller and more variable once the intervention becomes implemented more widely, and unanticipated consequences may begin to emerge. Long-term follow-up may be needed to determine whether short-term changes persist, and whether benefits inferred from surrogate outcomes in the original study do in fact occur. Although long-term follow-up of complex interventions is uncommon, such studies can be highly informative.[80] It is worth thinking about how to measure rare or long-term impacts, for example through routine data sources and record linkage, or by recontacting study participants. Plans for the collection of appropriate outcome data, and obtaining appropriate consents, should be built into the study design at the outset.

Part II Further questions

In part I we set out the key messages that researchers should bear in mind as they approach the evaluation of a complex intervention. In part II we list some further questions that evaluators might usefully ask themselves.

1. *Have you conducted a systematic review?* Systematic reviews of complex interventions can be problematic, as the methodology of how to find, review and combine data from complex intervention studies is not yet fully developed.[81] In particular, the methods for combining data from different study designs[82], and from differing variants of complex packages of care[83] present considerable problems. Heterogeneity is a different and more difficult type of problem in complex than in simple interventions, because of the wide scope for variation in the way interventions are delivered.[84]. Classifying the variant forms of a complex intervention, in terms of components, mode of delivery or intensity, in order to make sense of variability in outcomes, may not be straightforward.[85] It requires a theoretically informed understanding of the mechanisms of change underlying the intervention, and may be difficult to work out from the information available in published papers.

2. *Who is the intervention aimed at?* Complex interventions might be aimed at three different levels: at individuals (members of the public, patients, health or other practitioners, or policy makers); at community units (such as hospitals, schools or workplaces); at whole populations; or at more than one of these levels. If an intervention is seeking to achieve change at more than one level, e.g. influencing prescribing behaviour and patient outcomes, then processes and outcomes also need to be measured at each level. Changes in practice, such as the adoption of guidelines or implementation of service frameworks, are not necessarily mirrored by improvements in patient outcomes.[86]

3. *Can you describe the intervention fully?* A complex intervention, however 'complicated', should strive to be reproducible. This means that you need a full description of the intervention, and an understanding of its components, so that it can be delivered faithfully during the evaluation, allowing for any planned variation (see Question 4 below), and so that others can implement it outside your study. Given the length of time it can take to complete and write up an evaluation, it is useful to publish the study protocol (a description of the intervention and the evaluation design) while the evaluation is in progress. It is unlikely that a full description can be given in a scientific paper, so you should consider publishing a more detailed version on the web (see Question 9 below).

4. *How variable is the intervention – between sites, over time, etc?* An important but under-recognised aspect of a complex intervention is the fact that it may change with time, for pragmatic reasons, change of external factors, or as a result of learning during the evaluation. 'Fidelity' is not straightforward in relation to complex interventions.[87 88] In some evaluations, e.g. those comparing variants of a package of care,[89] or seeking to identify active ingredients within a complex intervention (see Case study 1), strict standardisation may be required to maintain clear separation between the different interventions, and to ensure that participants receive the right mix of ingredients. In such cases, controls must be put in place to limit unplanned variation.[90] But some interventions are deliberately designed to be adapted to local circumstances.[91 92] How much variation can be tolerated therefore depends on whether you are interested in efficacy or effectiveness[11]. Limiting variation in treatment may be desirable in an efficacy trial, but in a pragmatic, effectiveness study the statistical advantages and gain in 'internal validity' need to be weighed against the loss of generalisability or 'external validity'.[1] Investigators need to be clear about how much change or adaptation they would consider permissible.[2 93] Any variation in the intervention needs recording, whether or not it is intended, so that fidelity can be assessed in relation to the degree of standardisation required by the study protocol.[2]

5. *Can you describe the context and environment in which the evaluation is being undertaken?* Context is crucial: what works in one setting may not be as effective, or may even be harmful, elsewhere. The impact of a new intervention will depend on what provision already exists and interventions may have to be explicitly designed

to fit different contexts.[35] Sexual health interventions for low or middle income countries may be completely different to those appropriate to wealthier ones.[94] Furthermore, circumstances may change after the study has begun, for example through the introduction of some new policy, or, in the case of trials with long-term follow-up, more fundamental changes in the economic or political context. It is important therefore to develop a good understanding of the context in which the study is being carried out, and to monitor and document any significant changes.

6. *What user involvement is there going to be in the study?* Appropriate 'users' should be involved at all stages of the development, process and outcome analysis of a complex intervention, as this is likely to result in better, more relevant science and a higher chance of producing implementable data. A trial of the effect of providing home insulation on indoor environment, respiratory illness and self assessed health in New Zealand used a partnership research model in which agreements were signed with community groups who then worked closely with the research team (Case study 13). Qualitative research, as well as providing important insights into processes of change, can be a good way to involve users. It can complement user involvement in steering groups, and allows for a wider range of views to be canvassed and systematically incorporated into the design of an evaluation.[95 96]

7. *Is your study ethical?* Ethical problems of evaluating complex interventions need careful consideration. Investigators need to think about the ethics of their design in terms of the autonomy of participants and informed consent, and think through the possible effects of their trial in terms of effects on communities, possible adverse events, etc., in as robust a way as possible before seeking ethical approval. There may also be ethical objections to randomising in order to study health impacts, when the intervention has significant non-health benefits.[19] Ethical problems can, however, be overstated. Randomised consent designs (Case study 7) have attracted much debate, and are rarely used in practice, but are similar in key respects to cluster-randomised trials, a very widely accepted method in which consent to randomisation is not sought from individual patients.

8. *What arrangements will you put in place to monitor and oversee the evaluation?* Any complex intervention study should have appropriate data monitoring and steering committees[44] and comply with the relevant ethical and research governance frameworks. Appropriate in this context means proportionate to the various risks involved, such as risks to participants, financial risks, etc. A data monitoring committee is important in a clinical trial in which participants may be exposed to risk, or deprived of benefit, as a result of their allocation, but may be less relevant to an observational study in which participation incurs little or no additional risk. Whatever framework is adopted, it is important to incorporate an element of independent oversight, as difficult decisions may need to be taken about the direction and even continuation of the study.

9. *Have you reported your evaluation appropriately?* Where possible, evaluations should be reported according to established guidelines (Box 6), as this will help to ensure that the key information is available for replication studies, systematic reviews and guideline development. The CONSORT statement is a well-established standard for reporting randomised trials.[97] It has been extended to cover cluster-randomised trials[98] meta-analyses,[99] and non-pharmaceutical treatments.[100] Guidelines are also being developed for the reporting of non-randomised evaluations (the TREND statement[101]), for observational epidemiology more generally (the STROBE statement[102]), and for qualitative research.[82 103 104]

Box 6 Reporting guidelines

CONSORT Statement for the transparent reporting of clinical trials
http://www.consort-statement.org/?o=1001

STROBE Statement for strengthening the reporting of observational studies in epidemiology
http://www.strobe-statement.org/

TREND Statement for the transparent reporting of evaluations with non-randomised designs
http://www.trend-statement.org/asp/trend.asp

EQUATOR – promotes transparent and accurate reporting of health research
http://www.equator-network.org/?o=1001

10. Nevertheless, reporting of the evaluation of complex interventions remains poor in many cases.[105] A common failing is an inadequate description of the intervention[12] and work is under way to standardise the classification and description of complex interventions.[85 106] Use of graphical methods can help clarify what the interventions comprise, and highlight the key differences between experimental and control interventions.[107] Journals' space constraints have often prevented publication of a detailed description along with a report of the findings, but an increasing number provide for the publication of supplementary material on the web, and some, such as *Addiction*, now require a detailed protocol to be made available as a condition of publication. As suggested above (Question 3) prior publication of a protocol[26] methodological papers,[20] or an account of the development of the intervention[25] should also be considered (Case study 14). For a reliable assessment of the results of your evaluation, they need to be placed back in the context of other studies of similar interventions, if these exist.[108] For small studies, which may not produce statistically significant results if the initial assumptions about effect sizes, recruitment rates, etc., were over-optimistic, pooling makes the results far more useful.

11. *How will you analyse the data?* Because complex interventions typically involve multi-dimensional outcomes, they present a range of options for analysis, so statistical advice should be sought at the earliest opportunity. The more complex the intervention, the stronger the case for analysing data from observational and/or pilot studies first so as to understand the nature of the interactions amongst the various inputs and outcome measures before finalising the protocol for the main evaluation. The choice of analytic strategy will depend on the study design.[109-111] When the outcome measures can be combined into a small number of generally accepted summary measures, simple, robust, design-based methods of inference may be sufficient. For more complex outcome measures, model-based methods of inference may extract more useful information from the data, typically at the cost of requiring stronger assumptions. Options for dealing with complex patterns of statistical dependence include: multivariate approaches, such as principal components analysis,[112] which seek to reduce the dimensionality of a set of data; time series,[113] longitudinal and survival models[114] for analysing long and short sequences of measurements or time-to-event data respectively; multi-level models for hierarchically structured data[115] (e.g. patients within clinical teams, within wards, within hospitals) in which variation can occur at each level; and latent graphical modelling or structural equation modelling approaches[116] to characterising causal pathways among multiple, imperfectly measured input and outcome variables.

Part III Case studies

Case study 1

A causal modelling approach to developing a theory-based behaviour change intervention

Making explicit use of theory to develop an intervention prior to testing, and incorporating insights from the theory into an explicit model of how the intervention might alter behaviour, or affect other links in the causal chain between intervention and outcome, may lead to better-developed interventions, and also to better-designed evaluations.

Hardeman et al.[25] describe how they combined psychological theory with information about the epidemiology of Type 2 diabetes to develop an intervention to encourage people at risk to be more physically active, and to identify suitable outcome measures for use in a randomised trial. First the epidemiological literature was reviewed to identify the health outcome and the major risk factors, define the target group, estimate the likely impact of achievable behavioural change on the health outcome, and on the intervening physiological and biochemical variables, and to identify precise and reliable measures of the behaviour targeted by the intervention. Next, the researchers drew on psychological theory to identify determinants of the target behaviour, define intervention points, select appropriate behaviour change techniques, and develop measures of change in the determinants of behaviour.

The methods used were systematic reviews, expert meetings with researchers and practitioners, focus groups and interviews with the target population, and a survey of attitudes towards increasing physical activity. The outcome of this work, plus a further piece of action research, was an intervention, ProActive, and a set of measures that could be used to evaluate its effectiveness and investigate the mechanism by which any effects were achieved.[26] This was supplemented by a process study in which intervention sessions were recorded, so that delivery and receipt could be coded according to the theoretical model.[117 118] Importantly, given the significant lag between behaviour and outcome, the model was designed to enable the impact on final disease outcome to be estimated from trial data on changes in energy expenditure at 12 month follow-up, and epidemiological data on the relationship between physical activity and diabetes incidence.

The authors note that their approach is one of several that have been recommended and used for developing and evaluating complex interventions. Although the intervention was subsequently shown to be ineffective,[119] the value of a causal modelling approach is that it makes explicit the choice of intervention points and associated measures along the causal pathway. This allows researchers to assess *why* interventions are effective or not, as well as *how* effective they are, and to have greater confidence in modelling long-term disease outcomes from shorter term changes in behaviour.

Case study 2

An economic modelling approach to developing a complex intervention evaluation study

Economic modelling can provide important information about the likely cost-effectiveness of an intervention to inform decisions about whether to undertake a full scale evaluation.

Eldridge et al[32] used data from a pilot study and other sources to assess the likely cost-effectiveness of an intervention to reduce falls-related injuries in elderly people. The intervention was based on previously published guidance and consisted of a system of assessment and referral that healthcare professionals working in GP surgeries, A&E departments and nursing homes could use to assess elderly patients' risk of falling, and refer those deemed to be at high risk to appropriate sources of help. The aim was to conduct a pragmatic cluster-randomised trial that would inform the development and co-ordination of falls prevention services at a primary care trust level.

The pilot took place over 6 months in one primary care group. Information from the study was combined with routine data on injuries, mortality, health service use and costs, in order to model the contribution of the different stages (assessment, referral and treatment) to the overall effect of the intervention, and to explore its cost-effectiveness. Two models were constructed. One was used to assess the likelihood that a fall would be prevented, if an elderly person was assessed, referred and treated. The second model was used to estimate the long-term impact of the intervention on a hypothetical cohort of elderly people.

The results of the modelling exercise suggested that the intervention would prevent few falls, largely because few people at high risk or low risk would be assessed, and that even though the assessment was relatively cheap, it was unlikely to be good value for money. Sensitivity analyses suggested that the only way to improve this significantly would be to assess a far greater proportion of elderly people. One motivation for the modelling exercise was the experience of other community-based interventions, which tended to show small effects and problems reaching their target population. In the event, the main trial was not funded, but the pilot and the modelling exercise provided useful insights into the likely impact of the intervention, and showed that substantial changes would be needed to produce a viable method of preventing falls in elderly people. Torgerson and Byford[30] describe a number of other studies where economic modelling was used to inform decisions about trial design, covering issues such as trial size, choice of intervention and outcome measures.

Case study 3

Using pilot and feasibility studies to develop interventions and design evaluations

In addition to the use of theory and evidence from systematic reviews, pilot and feasibility studies are an essential step in the development and testing of an intervention, prior to a large-scale evaluation.

Rudolf et al.[120] describe a pilot study carried out to develop WATCH-IT, a community-based service to help obese children and adolescents in disadvantaged neighbourhoods to lose weight, and reduce their risk of early-onset atherogenesis. WATCH-IT involved appointing health trainers to set up and run clinics based in sports and community centres. Children with BMI above the 98th centile were eligible for the programme which involved a mixture of individual appointments with a trainer, group activity sessions, and group parenting sessions. The programme was developed following surveys of referrals to an endocrinology clinic and community paediatric dietetic service, relevant professionals and young people in the community. These indicated that frequency, flexibility and accessibility were likely to be key characteristics of a successful service.

After a year of operation in two centres, the programme was modified and extended to a number of other centres, and a pilot study initiated with a view to developing proposals for a randomised trial. The pilot study had three elements: a process evaluation, which measured attendance and the amount of support provided by professional staff, and also gathered their views of the service; a qualitative study of children enrolled in the programme, their parents and grandparents; and measurement of change in BMI SD scores, psychological wellbeing and quality of life over the six months from entry to the programme. Results indicated high levels of attendance and retention in the programme, improvements in self-esteem, and reductions in BMI among a majority of those children who continued to attend.

The study's starting point was the general lack of good evidence for effective ways of tackling childhood obesity, and complete absence of evidence on methods that might be applicable to disadvantaged children. The pilot suggested that the intervention was feasible in settings accessible to these children and their families, that retention in the programme was good, and that outcomes were favourable – in short, a strong candidate for further evaluation in a randomised trial.

Power et al.[94] describe a feasibility study of an adolescent sexual health intervention in rural Zimbabwe. Their starting point was the observation that sexual health interventions may fail because they lack an adequate theoretical basis, cannot be delivered as planned, or are inappropriate to their setting. Agreement was reached to incorporate a sexual health component within the existing life skills curriculum. Teachers in four rural secondary schools were trained to deliver the education materials in weekly lessons for two terms. Qualitative data was collected from pupils, parents, teachers, education officials, community leaders and health care providers. Teachers were also asked to record how the lessons had worked. The study revealed problems with the content of the materials, and showed that the secondary school classroom was inappropriate place to deliver sexual health education in rural Zimbabwe, given cultural norms regarding talk about sex, the general style of classroom teaching in the country, and the often exploitative relations between teachers and pupils. This led to a major redesign of the programme, with new materials written in indigenous languages, to be delivered by school-leavers living full-time in the study communities rather than by teachers, and supplemented by a community-based programme aimed at improving parent-child communication.

Case study 4

Clinical trials of complex interventions

There are fewer trials of surgical than of pharmaceutical interventions, and those there are tend to be of poorer quality. Trials are sometimes dismissed as impossible or inappropriate in surgery, for a variety of historical, cultural and technical reasons. The technical difficulties include lack of patient or clinician equipoise, difficulty of blinding, small effects of incremental improvements in technique, variability in the intervention over time ('learning curve effects') or between surgeons, and the related problem of fidelity and quality control.[121] These affect many other kinds of complex intervention, and have been successfully solved both in surgery and in other fields.

Arthroscopic lavage or debridement is frequently performed in patients with osteoarthritis of the knee that fails to respond to medical treatment. Although improvement in symptoms is commonly reported, the effect of the operation on the disease is unknown. Moseley *et al.*[122] conducted a randomised, double blind, placebo-controlled trial to determine the efficacy of arthroscopy in relieving pain and improving function. 180 patients were randomised to arthroscopic lavage and debridement, lavage alone, or placebo – a sham surgical procedure. Patients signed a consent form saying that they knew they might receive no treatment for their arthritis. Those assigned to placebo were anaesthetised (with an intravenous tranquiliser and an opioid rather than the general anaesthetic given to patients in the treatment group) and a sham operation performed lasting the same length of time as a debridement. A single surgeon performed all the real and simulated procedures. All the patients were kept in hospital overnight, and given the same postoperative care by nurses blind to their allocation. Outcomes were assessed at seven time points, from two weeks to two years post-operation, by assessors blind to allocation. The primary endpoint was knee pain at two years. To improve sensitivity, a further three observed and two self-reported pain and physical function outcomes were measured. In the analysis, differences in outcome were tested first for superiority of the treatment, and then for equivalence using the minimal important difference method. 165 patients completed the trial.

The simulation was highly successful – 13.8% of patients in the placebo group and 13.2% of those in the other arms guessed that they had received a placebo treatment. There was no difference in pain between the three groups at any of the time points, and some evidence of poorer function in the debridement group, though the analysis of equivalence suggested no minimally important differences between the groups. The surgeon was highly experienced in the procedure, suggesting that the results are likely to be robust with respect to efficacy. Although the patients were almost all men, the results are likely to be generalisable to women as earlier studies suggest no difference in their response to arthroscopic treatments. The meticulous design and conduct of the trial inspire confidence in the findings and suggest strongly that the positive outcomes reported by the observational studies may be placebo effects. The study is a good example of the value of a well-designed and conducted trial in an area where observational studies are likely to be prone to significant bias.

Case study 5

A cluster-randomised trial of a health protection intervention

Health protection and improvement interventions are sometimes regarded as off-limits for randomised trials, but examples do exist of successful trials of non-clinical interventions, in housing,[123] (see also Box 5) and other areas of social policy.[124] Contamination of the control group, leading to biased estimates of effect size, is often cited as an obstacle, but cluster randomisation, widely used in health services research, is one solution.

Risk of death from fire is strongly patterned by social class, and higher in households without a smoke alarm. DiGuiseppi et al[125] carried out a cluster-randomised controlled trial of the impact of providing free smoke alarms on the incidence of fires and related injuries in a deprived inner city population. A later trial explored variation in the prevalence of working smoke alarms by type of alarm[41] and a qualitative study was carried out to explore barriers and facilitators to the use of smoke alarms.[67]

In the first trial, 40 electoral wards in two Inner London Boroughs were paired by deprivation scores, and one in each pair randomly selected to receive the intervention. In the intervention wards, 20050 smoke alarms were distributed to high-risk households, together with batteries, fittings, leaflets in a range of local languages, and the offer of free installation, with the aim of increasing the rate of alarm ownership from 47% to 72% (the national average). The primary outcome was injury related to fire leading to attendance at an emergency department, hospitalisation or death during the two years following the intervention. A planned subgroup analysis was conducted of injuries judged to be preventable by two assessors blind to intervention status. A secondary outcome was the prevalence of working smoke alarms.

The results showed no benefit in terms of total or preventable injuries, hospitalisation or deaths related to fire, and no increase in the prevalence of working smoke alarms. Although serious injuries declined much faster in control than in intervention wards, the lack of increase in the use of smoke alarms in either suggests that contamination is not the explanation. The authors concluded instead that 'simply giving alarms to poor urban households is unlikely to reduce injuries related to fire.'

In the second trial, households were randomised to have one of five different types of smoke alarm installed. Alarms were tested and inspected in unannounced follow-up visits 15 months after installation. This study found that although some types of alarm were more likely to be working than others, only 54% of households had a working alarm at follow-up. The qualitative study indicated that households found the alarms hard to maintain, or actively disabled them because of the nuisance caused by false alarms – despite regarding themselves as being at risk.

A previous observational study in the US had shown a large decrease in hospitalisation and death related to fires in a high risk area following distribution of free smoke alarms. Failure to replicate this result in a randomised trial suggests the earlier result may have been biased. Taken together the results of the two trials and the qualitative study have important implications for the design both of smoke alarms and programmes to increase prevalence of functioning alarms. They also show the value of a mixed methods approach, incorporating a well-designed trial and a qualitative study to explore counter-intuitive findings.

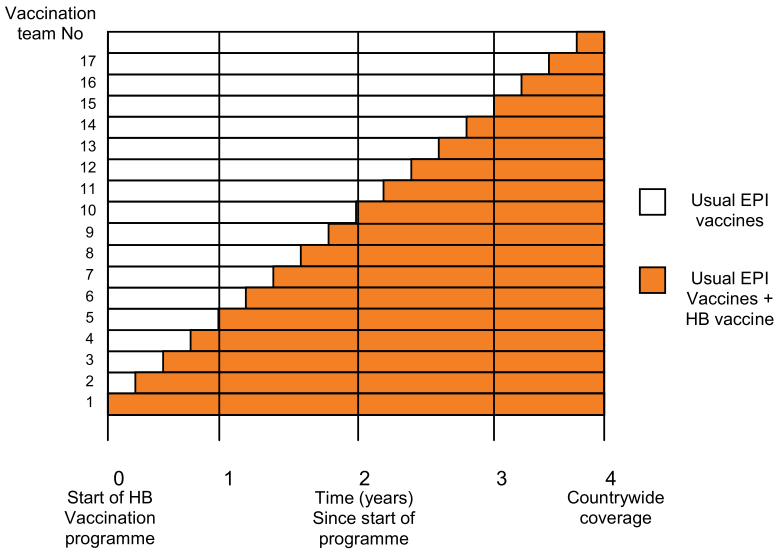
Case study 6

Stepped wedge designs – adapting implementation to enable evaluation

The stepped wedge design may be used as a solution to the ethical and practical problems of evaluating an intervention for which there is some evidence of effectiveness, or which, due to resource constraints or other reasons, cannot be made available to the whole population at once.[126] *It allows a randomised controlled trial to be conducted without delaying implementation of the intervention. Eventually, the whole population receives the intervention, but with randomisation built into the phasing of implementation.*

The Gambia Hepatitis Intervention Study (GHIS) was designed to evaluate the effect of Hepatitis B vaccination of newborn children on hepatocellular carcinoma and other liver diseases, in a country where nearly everyone was infected with HBV during childhood and one in five became chronic carriers.[127] Because liver disease only becomes common 20-30 years after infection, conducting a randomised trial before introducing a vaccination programme would have meant an unacceptable delay. Instead the evaluation was built into the implementation of the programme. At the time, the vaccine was expensive and in short supply, so vaccination of all newborn children in the Gambia was not feasible.

The existing vaccination programme in the Gambia was administered by 17 area-based teams. HBV vaccine was added to the vaccination schedule of each team in a random sequence, at 3 monthly intervals over a four-year period, until countrywide coverage was achieved (see figure).



Source[128]

To evaluate the effectiveness of the vaccine, the entire cohort of approximately 63,000 children who received the vaccines routinely available at the time (the white wedge in the figure) and 61,000 who also received the HBV vaccine (the dark wedge) are being followed up into adulthood using a system of cancer registration set up during the early stages of the vaccination programme, and enhanced for surveillance of cirrhosis and other forms of liver disease. More detailed surveillance was undertaken of a cohort of 1000 children from the first four areas to receive HBV vaccination. Their immune status was assessed at regular intervals into adolescence to determine whether revaccination on a mass basis was required. Cross-sectional surveys of unvaccinated children were also undertaken, and a number of ancillary studies built onto the basic framework of the GHIS.

Stepped wedge designs are rare (for a contemporary UK example see[129]). They have some drawbacks, such as greater complexity and a longer trial duration than parallel group studies, but also some important strengths, especially in overcoming objections to randomised studies of interventions which have already been shown to do more good than harm.[126]

Case study 7

Preference trials and other non-standard experimental designs

Practical or ethical obstacles to using randomised controlled trials can sometimes be overcome by the use of non-standard designs. Where patients have very strong preferences among treatments, basing treatment allocation on patients' preferences, or randomising patients before seeking consent, may be appropriate.

Where patients or other trial participants have strong preferences for treatment, they may refuse to take part in a trial, or drop out if not allocated to their favoured intervention, undermining the advantage of randomisation. If they do join and remain in the study, their preferences may bias the comparison of experimental and control treatments, by influencing their compliance with or response to treatment, or the way they report outcomes.[130] Although it is sometimes suggested that randomised trials are inappropriate where participants have strong preferences,[131] an alternative is to take variation in preferences into account in trial allocation, or in the analysis of outcomes. A number of ways of doing this have been developed.[132]

In the *comprehensive cohort design*, patients with strong preferences are offered their preferred intervention, while those without are randomised in the usual way. In the analysis, the randomised arms are compared first, then a further comparison is drawn with the preference arms, with adjustment for as many confounders as possible. In addition to helping overcome problems with recruitment, this design allows for an analysis of the effect of treatment preference on outcomes. A disadvantage is that deciding sample size is difficult if the distribution of preferences is unknown.

Two stage ('Wennberg') designs attempt to reduce baseline imbalances between randomised and preference groups, by first randomising participants to a group offered a choice of treatment, or to a group which is then randomised to treatment. In a variant of this approach (the *'Rucker' design*), participants in the first group who do not have strong preferences are also randomised to treatment.

Randomised consent ('Zelen') designs attempt to deal with preferences by first randomising participants to experimental and control groups, then seeking consent to treatment and follow-up from those in the experimental group. Patients in the control group receive standard care and are followed up unobtrusively. In the double consent variant, consent to treatment is sought from both groups and those who decline the allocated treatment receive the other.[133] This method has practical advantages in screening and other trials where blinding in the experimental group is impractical, but awareness of allocation among controls might lead to contamination, or where the process of recruitment and consent would constitute a partial intervention - for example, in a trial of sending postcards to patients to prevent readmission to hospital for deliberate self harm, where writing to seek consent from patients in the control group would mimic the intervention.[134] Randomised consent may also be ethically preferable to a conventional design in studies where it would be intolerable for those in the control group to know that a potentially life-saving treatment was being denied through randomisation.[130]

A systematic review[132] of preference trials using comprehensive cohort and two stage designs found that preferences affected recruitment but not retention. Preference and randomised groups differed in terms of baseline sociodemographic, health and clinical characteristics more often than would be expected by chance, though the only clear bias was an association between higher socio-economic position and having a treatment preference. Effects of preferences on outcomes were evident, though inconsistent in direction and not particularly associated with the use of subjective measures.

Preference trials based on comprehensive cohort designs have been used in a wide range of settings. Two stage designs are less widely used, possibly because they may reduce but not eliminate the problem of baseline imbalances.

The results of the review suggest the main advantage of preference designs is to improve recruitment, and therefore generalisability, though this needs to be offset against the need for larger sample sizes. If the key concern is to understand the relationship between preferences and outcomes, rather than to improve recruitment, an alternative is to measure preferences at baseline and then to analyse the interaction between preferences and outcomes in a fully randomised design.[135]

Zelen designs are infrequently used, and tend to be regarded as inappropriate for most therapeutic trials because of the incomplete consent, though there are circumstances in which they may be preferred on ethical or practical grounds.[130] In a variant of the method, Campbell et al[136] sought patients' consent to an observational study, then randomised them to intervention and control groups. Those in the control group were followed up as agreed, while the intervention group were asked to participate in a further study in which they received treatment. They conclude that for relatively safe interventions where there would be a risk of poor recruitment, retention or contamination in a conventional trial, 'the modified randomised consent design, in which patients first consent to an observational study, strikes a reasonable balance between the potentially competing imperatives of respect for patient autonomy and scientific rigour.'[136p 224]

Case study 8

Randomised n-of-1 designs

Parallel group RCTs aim to estimate the average effect of an intervention in a population, and provide little information about within or between person variability in response to interventions, or about the mechanisms by which effective interventions achieve change. N-of-1 trials, in which individuals undergo interventions with the order or scheduling decided at random, can be used to assess between and within person change, and to investigate theoretically predicted mediators of that change. The design allows intensive measurement over time, so that real-time change in cognitive, emotional and behavioural variables can be investigated.

There are two main functions for n-of-1 designs. The first is to evaluate the impact of interventions of proven effectiveness on individuals. This works best in chronic conditions, where the effects of alternative treatments 'wash-out' between treatment periods[137]- pain relief in chronic arthritis is a good example.

The second main function is to build the evidence of effectiveness:

Understanding heterogeneity: series of n-of-1 trials can be combined to provide an estimate of the overall effect of an intervention, and also estimates of the within and between person variability of the effect, which cannot be distinguished in the data from a conventional parallel group trial.[138]

Testing theory: n-of-1 trials can be used to test theory by experimentally manipulating postulated causal determinants of behaviour change. Similar methods can be used to identify the active components of an intervention.

Informing trial design: in the case of drug trials, n-of-1 studies can be used to identify which patients are most likely to respond, estimate required sample size, determine optimal dose, timing of onset and cessation of the drug's effect, and so on.[139] These arguments can be generalized to behaviour change and other non-drug interventions, with the methods adapted to take account of non-reversibility, e.g. by having multiple baseline measures and explicit criteria for changing between variants of the intervention.

Brookes et al[140] described a series of n-of-1 experiments carried out with patients with osteoarthritis of the knee to help design a complex intervention, involving multiple possible intervention components. In one set of experiments they tested a new type of knee brace, designed to provide local warmth as well as support. The comparison was with a traditional knee sleeve that provides the support but not the warmth. Some patients find warmth helpful, others seem to experience increased pain and distress, so the acceptability of using the new device, in comparison with the usual ones, needed to be assessed. The n-of-1 design was found to be highly acceptable to the patients, who gained more benefit from taking part and gaining more understanding of their symptoms and the factors that affect them, than they did from the interventions themselves.

An advantage of n-of-1 designs is that they satisfy the criteria of experimental randomization and internal validity without the large numbers of participants required by parallel group designs.

Case study 9

Natural experiments

Natural experiments offer opportunities for non-randomised evaluations of complex interventions.[5 6] They are useful in cases where deliberate manipulation of exposure is not possible, and work best in circumstances where a relatively large population is affected by a substantial change in a well-understood environmental exposure, and where exposures and outcomes can be captured through routine data sources, such as environmental monitoring and mortality records. They have been used effectively to measure the impact of controls on air pollution.

In 1990 the Hong Kong Government introduced legal restrictions on the use of fuel with a high sulphur content. The restriction was applied over one weekend and led to an immediate improvement in air quality and a fall in sulphur dioxide and airborne sulphate particulates. A number of studies were carried to assess the immediate and longer-term impact of the changes on respiratory health in children and on mortality. The mortality study[58] used data on deaths from respiratory and cardiovascular disease, cancer and other causes for the period 1985-95. Information on air pollutant concentrations over the period 1988-95 from Hong Kong's Environmental Protection Department was used to classify districts according to the extent of change in air quality. Results showed an immediate fall in seasonal deaths following introduction of the controls, followed by a seasonal peak, then a return to the original pattern 3-5 years post-introduction. The trend in deaths showed a decline in all-cause, respiratory and cardiovascular mortality after 1990, but not in deaths from other causes, and a greater reduction in areas with larger reductions in sulphur dioxide concentrations.

Also in 1990, the Irish Government implemented a ban on coal sales in the City of Dublin, following a period of increasing use of solid fuels for domestic heating during which episodes of increased air pollution were associated with increased in-hospital deaths. To study the effects of the ban, researchers compared air pollution, weather and cardiovascular, respiratory and all other non-trauma deaths for the six years before and six years following its introduction.[57] Analyses were stratified by season and adjusted for respiratory disease epidemics, weather, changes in the composition of Dublin's population, and secular changes in mortality in the rest of Ireland. There was an immediate, sustained and substantial fall in black smoke concentrations following the ban, and a more gradual fall in sulphur dioxide. Deaths from respiratory and cardiovascular disease fell markedly, as did all non-trauma deaths, though the decline was smaller.

Cohort studies[141] suggested a strong association between air pollution and mortality, though it was not known whether the excess deaths associated with episodes of air pollution represented a 'harvesting' effect, i.e. a bringing forward of deaths that were likely to have occurred soon in any case, or whether pollution controls would bring about a reduction in mortality. In the Hong Kong study, the availability of data on trends in deaths broken down by cause, and on the spatial distribution of changes in air pollution before and after the controls were introduced, enabled the researchers to distinguish immediate and longer term impacts of the intervention, and to attribute the decline in mortality to the introduction of the controls. Likewise, in the Dublin study, availability of trend data for both exposure and outcome, plus careful adjustment for potential confounders, enables confidence in attributing the decline in respiratory and cardiovascular deaths to the ban on coal sales.

Case study 10

Understanding processes

Clinical trials usually focus on outcomes, whereas evaluation of health promotion interventions are often more (and sometime exclusively) concerned with the process of effecting change. This dichotomy is unhelpful, as an understanding of process can provide useful insights into why an intervention achieves or fails to achieve the expected outcomes (see also Case study 5), but a process evaluation will be less informative if conducted without reference to outcomes, and may even be misleading if perceptions differ markedly from actual outcomes.

A trial[142] of educational outreach visits in which community pharmacists visited GP practices to encourage the use of prescribing guidelines for four commonly used interventions (aspirin as anti-platelet therapy, ACE inhibitors for heart failure, NSAIDs for osteoarthritic pain, and choice of antidepressant) found wide variation between practices and between guidelines in response to the visits. The primary outcome was change in proportion of patients treated according to the guidelines. Larger practices responded less than those with one or two practitioners. Across all four guidelines there was a net increase of 5% in the proportion of patients treated according to the guideline, but this varied from a 7% increase (for aspirin) to a 3% reduction for NSAIDs, with small increases for the remaining two.

To understand the variation in response, a post hoc evaluation[143] was conducted which explored the effect of the intervention on each step of a hypothesised pathway of change in GPs' prescribing behaviour. The pharmacists delivering the intervention collected information about practice-level participation in the study and GP attendance at the outreach visits, and completed semi-structured assessments of the visits. Two group discussions were held with the pharmacists, during and shortly after the visits, and a postal survey of GPs who had participated in at least one outreach session was carried out six months after the intervention was complete.

The results suggested that success in changing prescribing practice depended on a complex interaction between pharmacists, GPs and guideline topic. Pharmacists thought their confidence in the guidelines and in their own competence in delivering the intervention were important factors. They were satisfied with their own performance and reported good rapport with the GPs, but much lower expectations of change in prescribing behaviour. GPs rated the visits favourably and had good recall of the guideline recommendations, but this did not translate readily into improved prescribing. Some of the barriers were organisational, such as lack of practice systems for identifying patients with ischaemic heart disease (aspirin) or access to echocardiography diagnoses (ACE inhibitors). Others related to the nature of the guidelines: GPs preferred to implement recommendations based on clinical effectiveness (aspirin and ACE inhibitors) to those based on cost-saving (NSAIDs and antidepressants), where they were more sceptical of the evidence. They also tended to be reluctant to change prescribing in the face of patient resistance (NSAIDs) or where the change conflicted with local hospital policy (antidepressants). Although, as the authors point out, the ideal design would be a prospective rather than post hoc evaluation of the pathway to achieving behaviour change, this study, by linking process and outcome data together, provides valuable insights into the size and variability of the effect of the intervention on prescribing behaviour.

An example of such a prospective process evaluation to explain trial outcomes is Eccles et al's study[144] of the implementation of evidence-based guidelines for the management of adult patients with asthma or angina, using a computerised decision support system (CDSS). A cluster randomised controlled trial evaluation in 60 general practices in northern England) showed that there were no significant effects of CDSS on consultation rates, process of care measures (including prescribing) or any quality-of-life domain for either condition. A usage log across all practices showed that levels of use of the CDSS were low. Interviews conducted with physicians in five participating practices provided insights into why.[145] Interviewees were largely enthusiastic about the benefits of

computing for general practice and were optimistic about the potential for computers to present guidelines in a manageable format. However, the CDSS was felt by most practitioners to be difficult to use and unhelpful clinically. They believed that they were already familiar with the content of the guidelines, although they did not always follow recommendations for reasons that included limitations of the guidelines, patient preferences, lack of incentives and perceived structural barriers. The investigators concluded that even if it is possible to solve the technical hardware and software problems of producing a system that fully supports chronic disease management, there remains the challenge of integrating CDSS into clinical encounters in which busy practitioners manage patients with complex, multiple conditions.

Case study 11

Economic evaluation and complex interventions

Economic evaluation has an important role in both the planning and conduct of evaluations. Complex interventions are often expensive to evaluate, let alone implement, so economic considerations should inform the choice of intervention to evaluate and the detailed design of the evaluation. An economic evaluation conducted alongside an assessment of effectiveness will make the results of the evaluation as a whole much more useful for decision-making.

Research priorities are conventionally set on the basis of burden of disease, or some other assessment of the 'size of the problem'.[30] However, the largest problems are not necessarily associated with the most effective solutions, so focusing on them may not be the best use of resources.[146] Likewise, individual research funding decisions, and researchers' choices about which studies to undertake, can benefit from more explicit consideration of the value of the information that further research would generate. A formal framework has been developed for doing this which takes into account the costs of making the wrong decision on the basis of existing evidence.[29]

Once a decision has been made to evaluate an intervention, economic modelling can inform the design of the evaluation. For example, information about the extra cost of a new intervention can be used to work out how much more effective it would have to be in order to be cost-effective, and hence the sample size that would be needed to detect the relevant difference.[31] On the other hand, modelling may suggest that the experimental intervention is so unlikely to be cost-effective that a full trial was not needed (Case study 2).

There are a number of other issues that need to be taken into account in the design of the economic component of an evaluation:[147]

Appropriate choice of comparator: comparing the new intervention with standard treatment, if there is one, is more informative than comparing it to placebo. However, the trial may have to be larger as the difference between two active interventions is likely to be smaller than the difference between an active intervention and a placebo.

The intervention itself should be clearly defined, so that all relevant resource use can be identified

The primary outcome should be clearly identified. Where interventions have a diverse range of outcomes, it is preferable to combine them into a single generic measure such as the Quality Adjusted Life Year (QALY). Such measures also enable comparisons across cost-effectiveness studies.

Resource use and outcomes should be consistently recorded from baseline across an appropriate time period – i.e. the whole period across which costs and effects are expected to differ between interventions. If this exceeds the study follow-up period, the longer-term differences may have to be modelled.

Economic evaluation very often involves combining the results of an individual evaluation with external data, e.g. on service use or unit costs, or on costs and outcomes beyond the trial follow-up period. Record linkage can be useful, but needs to be carefully planned, to ensure that appropriate consent is obtained.

The perspective of the analysis should be explicit. A societal perspective, which takes account of the whole range of costs and effects is preferable to a narrower, e.g. health service, perspective.

To address these successfully, it is best to involve health economists early in the planning and design of the evaluation, so that the economic component is fully integrated.[148]

Case study 12

Implementation research: evidence in practice

Trials designed to provide robust estimates of treatment effectiveness may not provide accurate estimates of effectiveness once the treatment has passed into routine use, for example because they use highly selected patient populations or a restricted range of settings. Implementation research seeks to identify which techniques are effective for encouraging the translation of evidence into practice, and to provide information about 'real world' variability in effectiveness and cost effectiveness of interventions, and about the practicalities of introducing and sustaining new treatments or services. Case study 10 included two examples of implementation research. Two further examples are set out below.

Telephone consultations have been shown in a randomised controlled trial to achieve higher rates of participation in routine asthma reviews than the traditional method of face-to-face reviews.[149] But a large proportion of patients declined to take part in the trial, calling into question the generalisability of the findings to routine practice. In a subsequent implementation study,[150] Pinnock and colleagues compared three forms of asthma review service: a structured recall system in which patients were contacted up to three times by post or by memo issued with repeat prescriptions, and offered the choice of a telephone or a face-to-face review, then followed up opportunistically if they did not make an appointment; a similar recall system (but with no opportunistic follow-up) in which patients were offered only a face-to-face review; and usual care, in which there was no systematic recall.

Review rates increased over the baseline rates in all groups (national targets for reviewing asthma patients were introduced while the study was in progress), but increased most in the telephone-option group, even though a majority of patients in this group chose a face-to-face review. There was no clinical disadvantage associated with any of the review procedures, and the telephone-option procedure was cost-effective despite the extra cost of the opportunistic follow-ups. The study usefully extends the findings of the previous trial, by showing that a telephone option can increase participation in reviews, and how this can be implemented.

Evidence from trials suggests that nicotine replacement therapy (NRT) and bupropion (Zyban) help smokers to quit, and bupropion is slightly more effective in helping smokers to achieve long term abstinence.[151 152] An implementation study in two NHS Stop Smoking Services recorded the medication chosen by 2626 clients over a two year period, and compared this with a validated measure of abstinence at 3-4 weeks.[153 154] Information on service use and demographic characteristics was also collected. At 3-4 weeks, 34% of those using bupropion were abstinent, compared to 42% of NRT users (P=.003). The difference was not explained by demographic or treatment differences. Clients received similar levels of behavioural support, regardless of medication, and bupropion users tended to have characteristics, such as lower levels of dependence and greater likelihood of being in full-time employment, that favoured success. An alternative explanation lies with how clients obtained their medications.

Clients obtained NRT via a voucher redeemable at their local community pharmacy, or a letter to take to their GP to obtain a prescription, usually without requiring an appointment. Obtaining bupropion was more complicated. After a quit date had been set, the Stop Smoking Service would write to the client's GP recommending a prescription. The client would then need to book and attend an appointment in order to obtain their medication. Clients were expected to begin taking bupropion 1-2 weeks prior to the quit date, to allow time for a steady state concentration to be reached, but it is likely that many clients will not have reached this stage before quitting. Additionally, bupropion was only recommended for prescription 'with behavioural support', and repeat prescriptions were dependent both on abstinence and on continued attendance at the Stop Smoking Service. The relative complexity and additional conditions attached to obtaining bupropion may explain why it appeared to be less effective in establishing abstinence.

This study is a useful example of how implementation issues may effect compliance with, and effectiveness of, medications in health care as opposed to research settings.

Case study 13

Reporting evaluations of complex interventions

Evaluations of complex interventions should be reported whether or not the intervention was 'successful' and in a way that enables the intervention to be reproduced or adapted for the purposes of further research or for larger scale implementation. It is important to provide a full description of the intervention, and often useful to report the process of developing and implementing the intervention, as well as the results of the evaluation.

SHARE (Sexual Health and Relationships Education – Safe, Happy and Responsible) is a sex-education programme for 13-15 year old school pupils, aiming to reduce the incidence of unsafe sex and unwanted pregnancies and to improve the quality of young people's sexual relationships. It involves a training course for teachers, and a package of lessons for pupils in their third and fourth year of secondary school. It was developed against a background of worsening trends in key aspects of the sexual health of young people, and a history of poorly developed interventions lacking a clear conceptual basis. The results of evaluations were inconclusive, with stronger designs associated with smaller effects. Accordingly, SHARE was designed with an explicit theoretical basis and evaluated in a cluster-randomised trial comparing the new intervention with conventional sex education. The development of the intervention,[155 156] process of implementation[157] and outcomes[158 159] of the trial have all been reported comprehensively.

Development: the programme was developed on the basis of a wide-ranging review of relevant social science theory, previous evaluations, especially of interventions developed in the context of HIV prevention, and current practice in school sex education. It was adapted following each of two pilot studies and the intervention used in the trial represented a compromise between what pupils, teachers and schools could accept and accommodate, what previous research suggested was most likely to be effective, and the requirements of a randomised evaluation.[155]

Implementation: a study of the implementation of SHARE in the course of the trial found that, despite the efforts made to provide an intervention that could be accommodated within the school curriculum, and to encourage teachers to deliver the package in a standardised way, there were variations within and between schools in the way the package was delivered. Reasons identified included pressure on classroom time, competing initiatives, teachers' unfamiliarity with some of skills required and a sense of professional autonomy.[157]

Outcomes: the trial found that SHARE had no effect on sexual activity or risk-taking by age 16, or on conceptions or terminations of pregnancy by age 20, compared with conventional sex education, though pupils in the intervention schools had greater practical knowledge of sexual health and were less likely to regret their first sexual intercourse.[158 159]

The meticulous way in which SHARE was developed, implemented and tested means that despite its disappointing outcome, the evaluation has clear and important implications for both further research and policy. The results are unlikely to reflect failures of implementation or biases in the research, but real limits on the effectiveness of classroom delivered sex education.[158]

Case study 14

Involving users in the design and conduct of evaluations

Involving users in the design and conduct of evaluations, as well as being ethically preferable, has important practical advantages. Recruitment and retention are likely to be better if the intervention is valued by potential participants, concerns about fairness are addressed and, in the case of community-based interventions, community leaders support the evaluation. Involving users may also contribute to a better understanding of the process by which change is achieved.

The New Zealand Housing Insulation and Health Study[20 160] was a single-blind cluster randomised trial to determine whether fitting existing houses with a standard package of insulation measures improved indoor environments and reduced energy consumption, and improved occupants' health and wellbeing with a consequent reduction in their use of health care. Ethical approval was initially given for two pilot studies to test the feasibility of an experimental design. These led to decisions to insulate the control group participants' houses at the end of the study and to include a range of tenure types rather than focusing on social housing. Memoranda of understanding were signed with community organisations to make explicit the obligations on both sides and the fact that all participants would receive the intervention. These organisations were closely involved in conducting the study, for example by organising community meetings in which the study was explained to possible participants, recruiting local interviewers, and organising meetings to disseminate early results. Local health workers were employed to recruit participants into the study.

1350 households in seven predominantly low income communities were randomised. Baseline measures were taken before insulation was fitted in the intervention group, and both groups followed up one year later. Of 679 and 671 households randomised to the intervention and control groups, 563 and 565 respectively were retained in the study and included in the analysis. The results showed improvements in indoor environment in the intervention group and reductions in fuel use. Participants in intervention households reported better self-rated health, fewer symptoms and days off school or work, and fewer GP visits. There were similar numbers of hospital admissions for all causes in the intervention and control groups, though differences favouring the intervention group were slightly greater for respiratory than for other causes of admission. An economic evaluation indicated that, appropriately discounted, the health and energy benefits would substantially outweigh the costs of the intervention.

In a field where well-designed evaluations are rare and good quality randomised trials rarer still, the Housing Insulation and Health Study stands out as an example of good practice. It tested a relatively cheap and durable intervention that was targeted at people with respiratory illness in a low income community, and therefore had the potential to reduce health inequalities as well as to improve health. Because housing conditions are strongly correlated with income and other socio-economic risk factors for poor health, randomisation is an important safeguard against bias and confounding. Following pilot studies suggesting that drop-out might be high, the design was amended to ensure that all participants received the intervention either during or after the study. Community organisations were closely involved and a qualitative study[161] of this process was carried out. The study shows how involving communities in the design of an evaluation is not just compatible with the use of rigorous methods, but can also improve them.

Conclusions

The MRC framework has been extremely helpful to many researchers contemplating or executing the development, implementation or evaluation of complex interventions, and is widely cited in the international literature and in grant proposals. Since its publication much useful experience has been gained, and we therefore consider it timely to update and extend the guidelines. We have incorporated the key elements of this experience into this new guidance, either in the text, the references, or the case studies. We recognise that many of the issues that we have covered are still the subject of intense development and debate. We do not intend the guidance to be prescriptive, but rather to be helpful to researchers, research funders, policy-makers and other decision-makers in raising issues that they might address.

We have primarily aimed our messages at researchers. Perhaps the key message for research funders is the need for greater investment in developmental studies prior to large scale evaluations, and in implementation research. These will help to ensure a better return on investment in evaluation studies. The key message for policy makers is the need to incorporate evaluation considerations in the implementation of new initiatives, and wherever possible to allow for an experimental or a high quality non-experimental approach to be taken to the evaluation of significant initiatives where there is uncertainty about their effectiveness. Finally, we not expect this new version to be the final word. Experience will continue to accumulate and should be kept under review so that the guidance can be kept up to date and, no doubt, extended still further.

Bibliography

1. Rifkin A. Randomised controlled trials and psychotherapy research. *American Journal of Psychiatry* 2007;164(1):7-8.
2. Hawe P, Shiell A, Riley T, Gold L. Methods for exploring intervention variation and local context within a cluster randomised community intervention trial. *Journal of Epidemiology and Community Health* 2004;58:788-93.
3. Rychetnik L, Frommer M, Hawe P, Shiell A. Criteria for evaluating evidence on public health interventions. *Journal of Epidemiology and Community Health* 2002(56):119-27.
4. Wolff N. Randomised trials of socially complex interventions: promise or peril? *Journal of Health Services Research and Policy* 2001;6(2):123-6.
5. Ogilvie D, Mitchell R, Mutrie N, Petticrew M, Platt S. Evaluating health effects of transport interventions: methodologic case study. *American Journal of Preventive Medicine* 2006;31(2):118-26.
6. Petticrew M, Cummins S, Ferrell C, Findlay A, Higgins C, Hoy C, et al. Natural experiments: an under-used tool for public health. *Public Health* 2005;119:751-7.
7. Victora CG, Habicht J-P, Bryce J. Evidence-based public health: moving beyond randomised trials. *American Journal of Public Health* 2004;94(3):400-5.
8. MRC. A framework for the development and evaluation of RCTs for complex interventions to improve health. London: Medical Research Council, 2000:18.
9. Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, et al. Framework for the design and evaluation of complex interventions to improve health. *British Medical Journal* 2000;321:694-696.
10. Dieppe P. MRC PHSRN Complex interventions workshop May 2006. Workshop report for the PHSRN and HSPHRB, 2006.
11. Haynes B. Can it work? Does it work? Is it worth it? The testing of healthcare interventions is evolving. *British Medical Journal* 1999;319:652-3.
12. Michie S, Abraham C. Interventions to change health behaviours: evidence-based or evidence-inspired? *Psychology and Health* 2004;19(1):29-49.
13. Campbell NC, Murray E, Darbyshire J, Emery J, Farmer A, Griffiths F, et al. Designing and evaluating complex interventions to improve health care. *British Medical Journal* 2007;334:455-9.
14. Rutter M. Is Sure Start an effective preventive intervention? *Child and Adolescent Mental Health* 2006.
15. Belsky J, Melhuish E, Barnes J, Leyland AH, Romaniuk H, National Evaluation of Sure Start Research Team. Effects of Sure Start local programmes on children and families: early findings from a quasi-experimental, cross sectional study. *British Medical Journal* 2006;332:1476-81.
16. Wilkinson P, French R, Kane R, Lachowycz K, Stephenson J, Grundy C, et al. Teenage Conceptions, abortions, and births in England, 1994-2003, and the national teenage pregnancy strategy. *The Lancet* 2006;368:1879-86.
17. Creegan C, Hedges A. Towards a policy evaluation service: developing infrastructure to support the use of experimental and quasi-experimental methods. *Ministry of Justice Research Series*. London: Ministry of Justice, 2007.
18. Purdon S, Lessof C, Woodfield K, Bryson C. Research methods for policy evaluation. *Department for Work and Pensions Research Working Papers*. London: Department for Work and Pensions, 2001.
19. Thomson H, Hoskins R, Petticrew M, Ogilvie D, Craig N, Quinn T, et al. Evaluating the health effects of social interventions. *BMJ* 2004;328(7434):282-285.
20. Howden-Chapman P, Crane J, Matheson A, Viggers H, Cunningham M, Blakely T, et al. Retrofitting houses with insulation to reduce health inequalities: aims and methods of a clustered community-based trial. *Social Science and Medicine* 2005;61:2600-10.
21. Academy of Medical Sciences. Identifying the environmental causes of disease: how should we decide what to believe and when to take action. London: Academy of Medical Sciences, 2007.
22. Albarracin D, Gillette JC, Earl AN, Durantini MR, Moon-Ho H. A test of major assumptions about behaviour change: a comprehensive look at the effects of passive and active HIV-prevention interventions since the beginning of the epidemic. *Psychological Bulletin* 2005;131(6):856-97.
23. Noar SM, Zimmerman RS. Health behaviour theory and cumulative knowledge regarding health behaviours: are we moving in the right direction? *Health Education Research* 2005;20(3):275-90.
24. Michie S, Johnston M, Abraham C, Lawton R, Parker D, Walker A. Making psychological theory useful for implementing evidence-based practice: a consensus approach. *Quality and Safety in Healthcare* 2005;14:26-33.
25. Hardeman W, Sutton S, Griffin S, Johnston M, White A, Wareham NJ, et al. A causal modelling approach to the development of theory-based behaviour change programmes for trial evaluation. *Health Education Research* 2005;20(6):676-87.
26. Williams K, Prevost AT, Griffin S, Hardeman W, Hollingsworth W, Speigelhalter D, et al. The *ProActive* trial protocol - a randomised controlled trial of the efficacy of a family-based, domiciliary intervention programme to increase physical activity among individuals at high risk of diabetes. *BMC Public Health* 2004;4.

27. Eccles M, Johnston M, Hrisos S, Francis J, Grimshaw J, Steen N, et al. Translating clinicians' beliefs into implementation interventions (TRACII): a protocol for an intervention modelling experiment to change clinicians' intentions to implement evidence-based practice. *Implementation Science* 2007;2:27-32.
28. Bonetti D, Eccles M, Johnston M, Steen N, Grimshaw J, Baker R, et al. Guiding the design and selection of interventions to influence the implementation of evidence-based pactice: an experimental simulation of a complex intervention trial. *Social Science and Medicine* 2005;60:2135-2147.
29. Claxton K, Sculpher M, Drummond M. A rational framework for decision-making by the National Institute for Clinical Excellence. *Lancet* 2002;31:711-15.
30. Torgerson D, Byford S. Economic modelling before clinical trials. *British Medical Journal* 2002;325:98.
31. Torgerson D, Campbell M. Cost effectiveness calculations and sample size. *British Journal of General Practice* 2000;321:627.
32. Eldridge S, Spencer A, Cryer C, Pearsons S, Underwood M, Feder G. Why modelling a complex intervention is an important precursor to trial design: lessons from studying an intervention to reduce falls-related injuries in elderly people. *Journal of Health Services Research and Policy* 2005;10(3):133-42.
33. Collins LM, Murphy SA, Nair VN, Stretcher VJ. A strategy for optimizing and evaluating behavioral interventions. *Annals of Behavioral Medicine* 2005;30(1):65-73.
34. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *American Journal of Public Health* 1999;89:1322-7.
35. National Institute for Health and Clinical Excellence. Behaviour Change at Population, Community and Individual Levels. *NICE Public Health Guidance*. London: NICE, 2007.
36. Glasgow RE, Lichtenstein E, Marcus AC. Why don't we see more translation of health promotion research into practice? Rethinking the efficacy-to-effectiveness transition. *American Journal of Public Health* 2003;93(8):1261-7.
37. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of research for decision-making in clinical and health policy. *Journal of the American Medical Association* 2002;290(12):1624-32.
38. Eldridge S, Ashby D, Feder G, Rudnicka AR, Ukoumunne OC. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clinical Trials* 2004;1:80-90.
39. Scheel I, Hagen K, Oxman A. The unbearable lightness of healthcare policy-making: a description of a process aimed at giving it some weight. *Journal of Epidemiology and Community Health* 2003;57:483-87.
40. Armstrong D, Winder R, Wallis R. Impediments to policy implementation: the offer of free installation of central heating to an elderly community has limited uptake. *Public Health* 2006;120:121-6.
41. Rowland D, DiGuiseppe C, Roberts I, Curtis K, Roberts H, Ginnelly L, et al. Prevalence of working smoke alarms in local authority inner city housing: randomised controlled trial. *British Medical Journal* 2002;325:998-1001.
42. Bower P, Wilson S, Mathers N. Short report: How often do UK primary care trials face recruitment delays? *Family Practice* 2007.
43. McDonald A, Knight R, Campbell M, Entwistle V, Grant A, Cook J, et al. What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. *Trials* 2006;7(9).
44. Prescott R, Counsell C, Gillespie W, Grant A, Russell I, Kiauka S, et al. Factors that limit the quality, number and progress of randomised controlled trials. *Health Technology Assessment* 1999;3(20).
45. McKee M, Britton A, Black N, McPherson K, Sanderson C, Bain C. Interpreting the evidence: choosing between randomised and non-randomised studies. *British Medical Journal* 1999;319:312-5.
46. Walwyn R, Wessely S. RCTs in psychiatry: challenges and the future. *Epidemiologia e Pschiatria Sociale* 2005;14(3):127-31.
47. Collins R, MacMahon S. Reliable assessment of the effects of treatment on mortality and major morbidity. *Lancet* 2001;357:373-80.
48. Eccles M, Grimshaw J, Campbell M, Ramsay C. Research designs for studies evaluating the effectiveness of change and improvement strategies. *Quality and Safety in Healthcare* 2003;12:47-52.
49. Ukoumunne OC, Gulliford MC, Chinn S, Sterne JAC, Burney PGJ, Donner A. Methods in health service research: Evaluation of health interventions at area and organisation level. *BMJ* 1999;319(7206):376-379.
50. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;312(7040):1215-1218.
51. Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. *British Medical Journal* 2007;334:349-51.
52. Greenland S. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology* 2000;29:722-9.
53. D'Agostino R. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* 1998;17(19):2265-81.
54. Deeks J, Dinnes J, D'Amico R, Sowden A, Sakarovich C, Song F, et al. Evaluating non-randomised intervention studies. *Health Technology Assessment* 2003;7(27).

55. MacLehose R, Reeves B, Harvey I, Sheldon T, Russell I, Black A. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. *Health Technology Assessment* 2003;4(34).
56. Gunnell D, Fernando R, Hewagama M, Priyangika W, Konradsen F, Eddleston M. The impact of pesticide regulations on suicide in Sri Lanka. *International Journal of Epidemiology* 2007;36:1235-1242.
57. Clancy L, Goodman P, Sinclair H, Dockery DW. Effect of air pollution control on death rates in Dublin, Ireland: an intervention study. *Lancet* 2002;360:1210-14.
58. Hedley A, Wong C, Thach T, Ma S, Lam T, Anderson H. Cardiorespiratory and all-cause mortality after restrictions on sulphur content of fuel in Hong Kong: an intervention study. *Lancet* 2002;360(9346):1646-52.
59. Akhtar PC, Currie DB, Currie CE, Haw SJ. Changes in child exposure to environmental tobacco smoke (CHETS) study after implementation of smoke-free legislation in Scotland: national cross sectional survey. *British Medical Journal* 2007;335:545-9.
60. Haw SJ, Gruer L. Changes in exposure of adult non-smokers to secondhand smoke after implementation of smoke-free legislation in Scotland: national cross sectional survey. *British Medical Journal* 2007;335:549-52.
61. Semple S, Creely K, Naji A, Miller B, Ayres J. Second hand smoke levels in Scottish pubs: the effect of the smoke-free legislation. *Tobacco Control* 2007;16:127-32.
62. MacMahon S, Collins R. Reliable assessment of the effects of treatment on mortality and major morbidity, II: observational studies *The Lancet* 2001;357:455-62.
63. Vandenbroucke JP. When are observational studies as credible as randomised trials? *The Lancet* 2004;363:1728-31.
64. Blair PS, Fleming PJ, Smith IJ, Platt MW, Young J, Nadin P, et al. Babies sleeping with parents: case-control study of factors influencing the risk of the sudden infant death syndrome • Commentary: Cot death---the story so far. *BMJ* 1999;319(7223):1457-1462.
65. Fleming PJ, Blair PS, Bacon C, Bensley D, Smith I, Taylor E, et al. Environment of infants during sleep and risk of the sudden infant death syndrome: results of 1993-5 case-control study for confidential inquiry into stillbirths and deaths in infancy. *BMJ* 1996;313(7051):191-195.
66. Gilbert R, Salanti G, Harden M, See S. Infant sleeping position and the sudden infant death syndrome: systematic review of observational studies and historical review of recommendations from 1940 tom 2002. *International Journal of Epidemiology* 2005;34:874-87.
67. Roberts H, Curtis K, Liabo K, Rowland D, DiGuiseppe C, Roberts I. Putting public health evidence into practice: increasing the prevalence of working smoke alarms in disadvantaged inner city housing. *Journal of Epidemiology and Community Health* 2004;854:280-85.
68. Bonell C, Oakley A, Hargreaves J, Strange V, Rees R. Assessment of generalisability in trials of health interventions: suggested framework and systematic review. *British Medical Journal* 2006;333:346-9.
69. Oakley A, Strange V, Bonell C, Allen E, Stephenson J, RIPPLE Study Team. Process evaluation in randomised controlled trials of complex interventions. *British Medical Journal* 2006;332:413-6.
70. Roen K, Arai L, Roberts H, Popay J. Extending systematic reviews to include evidence of implementation: methodological work on a review of community-based initiatives to prevent injuries. *Social Science and Medicine* 2006;63:1060-71.
71. Briggs A. Handling uncertainty in economic evaluation. *British Medical Journal* 1999;319:120.
72. Briggs A. Economic evaluation and clinical trials: size matters. *British Medical Journal* 2000;321:1362-3.
73. Oxman A, Thomson M, Davis D, Haynes R. No magic bullets: a systematic review of 102 trials of interventions to improve professional practice. *Canadian Medical Association Journal* 1995;153(10):1423-31.
74. NHS Centre for Reviews and Dissemination. Getting evidence into practice. *Effective Healthcare* 1999;5(1).
75. Grimshaw J, Thomas R, MacLennan G, Fraser C, Ramsay C, Vale L, et al. Effectiveness and efficiency of guideline dissemination and implementation strategies. *Health Technology Assessment* 2004;8(6).
76. Glasgow RE, Davidson KW, Dobkin PL, Ockene J, Spring B. Practical behavioural trials to advance evidence-based behavioral medicine. *Annals of Behavioral Medicine* 2006;31(1):5-13.
77. Lavis J, Davies H, Oxman A, Denis J-L, Golden-Biddle K, Ferlie E. Towards systematic reviews that inform healthcare management and policy-making. *Journal of Health Services Research and Policy* 2005;10(Suppl 1):35-48.
78. Michie S, Johnston M. Changing clinical behaviour by making guidelines specific. *British Medical Journal* 2004;328:343-5.
79. Bero LA, Grilli R, Grimshaw J, Harvey E, Oxman AD, Thomson MA. Closing the gap between research and practice: an overview of systematic reviews of interventions to promote the implementation of research findings. *British Medical Journal* 1998;317:465-8.
80. Wortman PM. An exemplary evaluation of a program that worked: The High/Scope Perry Preschool Project. *American Journal of Evaluation* 1995;16(3):257-65.
81. Petticrew M. Why certain systematic reviews reach uncertain conclusions. *British Medical Journal* 2003;326:756-8.
82. Dixon-Woods M, Fitzpatrick R. Qualitative research in systematic reviews. *British Medical Journal* 2001;323:765-6.

83. Caldwell DM, Ades A, Higgins J. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *British Medical Journal* 2005;331:879-900.

84. Herbert RD, Bo K. Analysis of quality of interventions in systematic reviews. *British Medical Journal* 2005;331:507-9.

85. Michie S, Abraham C. Developing a taxonomy of behaviour change intervention techniques. *Psychology and Health* 2007;22(Suppl 1).

86. Moher M, Yudkin P, Wright L, Turner R, Fuller A, Schofield T, et al. Cluster-randomised controlled trial to compare three methods of promoting secondary prevention of coronary heart disease in primary care. *British Medical Journal* 2001;322:1-7.

87. Bellg AJ, Borrelli B, Resnick B, Hecht J, Sharp Minicuchi D, Ory M, et al. Enhancing treatment fidelity in health behaviour change studies: best practices and recommendations from the NIH Beviour Change Consortium. *Health Psychology* 2004;23(5):543-51.

88. Leventhal H, Friedman MA. Does establishing fidelity of treatment help in understanding treatment efficacy? Comment on Bellg et al. (2004). *Health Psychology* 2004;23(5):452-6.

89. Sharp DM, Power KG, Swanson V. Reducing therapist contact time in CBT for panic disorder and agoraphobia and in primary care: global measures of outcome in a randomised controlled trial. *British Journal of General Practice* 2000;50:963-968.

90. Farmer A, Wade A, Goyder E, Yudkin P, French D, Craven A, et al. Impact of self-monitoring of blood glucose in the management of patients with non-insulin treated diabetes: open parallel group randomised trial. *British Medical Journal* 2007;335:132-9.

91. Patton G, Bond L, Butler H, Glover S. Changing schools, changing health? Design and implementation of the Gatehouse Project. *Journal of Adolescent Health* 2003;33:231-39.

92. Patton GC, Bond L, Carlin JB, Thomas L, Butler H, Glover S, et al. Promoting social inclusion on schools: a group-randomized trial of effects on student health risk behaviour and well-being. *American Journal of Public Health* 2006;96(9):1582-7.

93. Hawe P, Shiell A, Riley T. Complex interventions: how "out of control" can a randomised trial be? *British Medical Journal* 2004;328:1561-63.

94. Power R, Langhaug L, Nyamurera T, Wilson D, Bassett M, Cowan F. Developing complex interventions for rigorous evaluation - a case study from rural Zimbabwe. *Health Education Research* 2004;19(5):570-575.

95. Yardley L, Bishop FL, Beyer N, Hauer K, Kempen GIJM, Piot-Ziegler C, et al. Older People's Views of Falls-Prevention Interventions in Six European Countries. *Gerontologist* 2006;46(5):650-660.

96. Yardley L, Donovan-Hall M, Francis K, Todd C. Older people's views of advice about falls prevention: a qualitative study. *Health Educ. Res.* 2006;21(4):508-517.

97. Moher D, Schulz KF, Altman DG. Revised recommendations for improving the quality of reports of parallel group randomized trials 2001. *The Lancet* 2001;357:1191-1194.

98. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster-randomised trials. *British Medical Journal* 2004;328:702-8.

99. Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUORUM statement. *Lancet* 1999;354:1896-900.

100. Boutron I, Moher D, Altman D, Scultz K, Ravaud P. Extending the CONSORT Statement to Randomized Trials of Non-pharmacologic Treatment: Explanation and Elaboration. *Annals of Internal Medicine* 2008;148:295-309.

101. DesJarlais DC, Lyles C. Improving the reporting quality of nonrandomized evaluations of behavioural and public health interventions: the TREND statement. *American Journal of Public Health* 2004;94(3):361-6.

102. von Elm E, Altman D, Egger M, Pocock SJ, Gotzsche P, Vandenbroucke JP, et al. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *British Medical Journal* 2007;335:335-8.

103. Dixon-Woods M, Shaw R, Agarwal S. The problem of appraising qualitative research. *Quality and Safety in Healthcare* 2004;13:223-5.

104. Dixon-Woods M, Sutton A, Shaw R, Miller T, Smith J, Young B, et al. Appraising qualitative research for inclusion in systematic reviews: a quantitative and qualitative comparison of three methods. *Journal of Health Services Research and Policy* 2007;12(1):42-7.

105. Dumbrowski SU, Sniehotta FF, Avenell AA, Coyne JC. Towards a cumulative science of behaviour change: do current conduct and reporting of behavioural interventions fall short of best practice? *Psychology and Health* 2007;22(8):869-74.

106. Michie S, Johnston M, Francis J, Hardeman W, Eccles M. From theory to intervention: mapping theoretically derived behavioural determinants to behaviour change techniques. *Applied Psychology: an International Review* (in press).

107. Perera R, Heneghan C, Yudkin P. Graphical method for depicting randomised trials of complex interventions. *British Medical Journal* 2007;334:127-9.

108. Wilson P, Petticrew M. Knowledge transfer more harm than good: why do we promote the findings of single research studies? *British Medical Journal* Forthcoming.

109. Breslow N, Day N. Statistical Methods in Cancer Research: The Analysis of Case–Control Studies *IARC Scientific Publications* 1980;32:5-388.

110. Breslow N, Day N. Statistical Methods in Cancer Research: The Design and Analysis of Cohort Studies. *IARC Sci Publ.* 1987;82:1-406.

111. Ukoumunne OC, Gulliford MC, Chinn S, Sterne JAC, Burney PGJ. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technology Assessment* 1999;3(5).

112. Manly B. *Multivariate Statistical Methods: A primer*. 3 ed. Boca Raton: Chapman and Hall/CRC, 2005.

113. Chatfield C. *The Analysis of Time Series: An Introduction*. 6 ed. Boca Raton: Chapman and Hall/CRC, 2004.

114. Diggle P, Heagerty P, Liang K-Y, Zeger S. *Analysis of Longitudinal Data*. 2 ed. Oxford: Oxford University Press, 2002.

115. Goldstein H, . *Multilevel Statistical Models*. 3 ed. London: Hodder Arnold, 2003.

116. Skrondal A, Rabe-Hesketh S. *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Boca Raton: Chapman and Hall/CRC, 2004.

117. Hardeman W, Michie S, Fanshawe T, Prevost AT, McLoughlin K, Kinmonth AL. Fidelity of delivery of a physical activity intervention. Predictors and consequences. *Psychology and Health* In press.

118. Michie S, Hardeman W, Fanshawe T, Prevost AT, Taylor L, Kinmonth AL. Investigating theoretical explanations for behaviour change. The case study of ProActive. *Psychology and Health* In press.

119. Kinmonth A-L, Wareham NJ, Hardeman W, Sutton S, Prevost AT, Fanshawe T, et al. Efficacy of a theory-based behavioural intervention to increase physical activity in an at-risk group in primary care (ProActive UK): a randomised trial. *The Lancet* 2008;371:41-8.

120. Rudolf M, Christie D, McElhone S, Sahota P, Dixey R, Walker J, et al. WATCH IT: a community based programme for obese children and adolescents. *Archives of Disease in Childhood* 2006;91:736-739.

121. McCulloch P, Taylor I, Sasako M, Lovett B, Griffin D. Randomised trials in surgery: problems and possible solutions. *British Medical Journal* 2002;324:1448-51.

122. Moseley J, O'Malley K, Petersen N, Menke T, Brody B, Kuykendall D, et al. A controlled trial of arthroscopic surgery for osteoarthritis of the knee. *New England Journal of Medicine* 2002;347(2):81-8.

123. Krieger J, Takaro T, Song L, Weaver M. The Seattle-King County Healthy Homes Project: a randomised controlled trial of a community health worker intervention to decrease exposure to indoor asthma triggers. *American Journal of Public Health* 2005;954:652-59.

124. Hutchings J, Gardner F, Bywater T, Daley D, Whitaker C, Jones K, et al. Parenting intervention in Sure Start services for children at risk of developing conduct disorder: pragmatic randomised controlled trial. *British Medical Journal* 2007;334:678-82.

125. diGuiseppe C, Roberts I, Wade A, Sculpher M, Edwards P, Godward C, et al. Incidence of fires and related injuries after giving out free smoke alarms: cluster randomised controlled trial. *British Medical Journal* 2002;325:995-8.

126. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Medical Research Methodology* 2006;6:54-63.

127. Hall A, Inskip H, Loik F, Day N, The Gambia Hepatitis Study Group. The Gambia Hepatitis Intervention Study. *Cancer Research* 1987;47:5782-5787.

128. Hainaut P. The Gambia Hepatitis Intervention Study: Report to the experts of the IARC Scientific Council Site Review, January 2004, 2004.

129. Stone S, Slade R, Fuller C, Charlett A, Cookson B, Teare L, et al. Early communication: Does a national campaign to improve hand hygiene in the NHS work? Initial English and Welsh experience from the NOSEC study (National Observational Study to Evaluate the CleanYourHands Campaign). *Journal of Hospital Infection* 2007;66(3):293-6.

130. Torgerson D, Roland M. Understanding controlled trials: what is Zelen's design? *British Medical Journal* 1998;316:606.

131. McPherson K, Chalmers I. Incorporating patient preferences into clinical trials - information about patients' preference must be obtained first. *British Medical Journal* 1998;317:78.

132. King M, Nazareth I, Lampe F, Bower P, Chandler M, Morou M, et al. Impact of participant and physician intervention preferences on randomised trials. *Journal of the American Medical Association* 2005;293(9):1089-1099.

133. Altman D, Whitehead J, Parmar M, Stenning S, Fayers P, Machin D. Randomised consent designs in cancer clinical trials. *European Journal of Cancer* 1995;31A(12):1934-44.

134. Carter G, Clover K, Whyte I, Dawson A, D'Este C. Postcards from the EDge project: randomised controlled trial of an intervention using postcards to prevent repetition of hospital treated deliberate self-poisoning. *British Medical Journal* 2005;331:805-10.

135. Klaber-Moffett J, Jackson D, Richmond S, Hahn S, Coulton S, Farrin A, et al. Randomised trial of a brief physiotherapy intervention compared with usual physiotherapy for neck pain patients: outcomes and patients' preference. *British Medical Journal* 2005;330:75-81.

136. Campbell R, Peters T, Grant C, Quilty B, Dieppe P. Adapting the randomized consent (Zelen) design for trials of behavioural interventions for chronic disease: feasibility study. *Journal of Health Services Research and Policy* 2005;10(4):220-225.

137. Guyatt G, Sackett D, Adachi J, Roberts R, Chong J, Rosenbloom D, et al. A clinician's guide for conducting randomised trials in individual patients. *Canadian Medical Association Journal* 1988;139(6):497-503.

138. Zucker D, Schmid C, McIntosh M, D'Agostino R, Selker H, Lau J. Combining single patient (n-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. *Journal of Clinical Epidemiology* 1997;50(4):401-10.

139. Guyatt G, Heyting A, Jaeschke R, Keller J, Adachi J, Roberts R. N-of-1 randomised trials for investigating new drugs. *Controlled Clinical Trials* 1990;11:88-100.

140. Brookes ST, Biddle L, Paterson C, Woolhead G, Dieppe P. "Me's me and you's you": exploring patients' perspectives of single patient (n-of-1) trials in the UK. *Trials* 2007;8:10-18.

141. Hoek G, Brunekeef B, Goldbohm S, Fischer P, Brandt Pvd. Association between mortality and indicators of traffic related air pollution in the Netherlands: a cohort study. *Lancet* 2002;360:1203-9.

142. Freemantle N, Nazareth I, Eccles M, Wood J, Haines A, EBOR Triallists. A randomised controlled trial of the effect of educational outreach by community pharmacists on prescribing in UK general practice. *British Journal of General Practice* 2002;52:290-5.

143. Nazareth I, Freemantle N, Duggan C, Mason J, Haines A. Evaluation of a complex intervention for changing professional behaviour: the Evidence Based Out Reach (EBOR) Trial. *Journal of Health Services Research and Policy* 2002;7(4):230-8.

144. Eccles M, McColl E, Steen N, Rousseau N, Grimshaw J, Parkin D, et al. Effect of computerised evidence based guidelines on management of asthma and angina in adults in primary care: cluster randomised controlled trial. *British Medical Journal* 2002;325:941-8.

145. Rousseau N, McColl E, Newton J, Grimshaw J, Eccles M. Practice based, longitudinal, qualitative interview study of computerised evidence based guidelines in primary care. *British Medical Journal* 2003;326:314-22.

146. Normand C. Ten popular health economic fallacies. *Journal of Public Health Medicine* 1998;20(2):129-32.

147. Byford S, Mcdaid D, Sefton T. *Because it's worth it. A practical guide to conducting economic evaluations in the social welfare field.* York: Joseph Rowntree Foundation, 2002.

148. Sefton T, Byford S, McDaid D, Hills J, Knapp M. *Making the most of it: economic evaluation in the social welfare field.* York: Joseph Rowntree Foundation, 2002.

149. Pinnock H, Bawden R, Proctor S. Accessibility, acceptability and effectiveness of telephone reviews for asthma in primary care: randomised controlled trial. *British Medical Journal* 2003;326:477-479.

150. Pinnock H, Adlem L, Gaskin S, Harris J, Snellgrove C, Sheikh A. Accessibility, clinical effectiveness and practice costs of providing a telephone option for routine asthma reviews: phase IV controlled implementation study. *British Journal of General Practice* 2007;57:714-22.

151. Hughes J, Stead L, Lancaster T. Antidepressants for smoking cessation. *Cochrane Database Syst Rev* 2007;(1):CD000031.

152. Silagy C, Lancaster T, Stead L, Mant D, Fowler G. Nicotine replacement therapy for smoking cessation. *Cochrane Database Syst Rev* 2004;(3):CD000146.

153. McEwen A, West R, McRobbie H. Do implementation issues influence the effectiveness of medications? The case of NRT and bupropion in NHS Stop Smoking Services. *Patient Education and Counselling.* Forthcoming.

154. West R, McNeill A, Raw M. Smoking cessation guidelines for health professionals: an update. *Thorax* 2000;55(12):987-999.

155. Wight D, Abraham C. From psychosocial theory to sustainable classroom practice: developing a research-based teacher-delivered sex education programme. *Health Education Research* 2000;15(1):25-38.

156. Wight D, Abraham C, Scott S. Towards a psychosocial theoretical framework for sexual health promotion. *Health Education Research* 1998;13(3):317-30.

157. Buston K, Wight D, Hart G, Scott S. Implementation of a teacher-delivered sex education programme: obstacles and facilitating factors. *Health Education Research* 2002;17(1):59-72.

158. Henderson M, Wight D, Raab GM, Abraham C, Parkes A, Scott S, et al. Impact of a theoretically based sex education programme (SHARE) delivered by teachers on NHS registered conceptions and terminations: final results of a cluster randomised trial. *British Medical Journal* 2006;334.

159. Wight D, Raab GM, Henderson M, Abraham C, Buston K, Hart G, et al. Limits of teacher delivered sex education: interim behavioural outcomes from randomised trial. *British Medical Journal* 2002;324:1430-6.

160. Howden-Chapman P, Matheson A, Crane J, Viggers H, Cunningham M, Blakely T, et al. Effect of insulating houses on health inequality: cluster randomised study in the community. *British Medical Journal* 2007.

161. Matheson A, Howden-Chapman P, Dew K. Engaging communities to reduce health inequalities: why partnership? *Social Policy Journal of New Zealand* 2005;26:1-16.