Comment

Evidence into practice, experimentation and quasi experimentation: are the methods up to the task?

N Freemantle, J Wood, F Crawford

Abstract

Objective—Methodological review of evaluations of interventions intended to help health professionals provide more effective and efficient health care, motivated by the current experience of NHS Research and Development in England. Emphasis upon the forms of research appropriate to different stages in the development and evaluation of interventions, the use of experimental and quasi experimental designs, the methods used in systematic reviews and meta analyses.

Method—A proposed development process is derived from that used in the development of drugs. The strengths and weaknesses of different experimental and quasi experimental designs are derived from published methodological literature and first principles. Examples are drawn from the literature.

Results—Like pharmaceuticals, implementation of interventions need to go through several stages of development before they are evaluated in designed experiments. Where there are practical reasons that make random allocation impossible in quantitative evaluations, quasi experimental methods may provide useful information, although these studies are open to bias. It is rare for a single study to provide a complete answer to important questions, and systematic reviews of all available studies should be undertaken. Meta analytic techniques go some way towards countering the low power of many existing studies, reduce the risk of bias, and avoid the subjective approaches that may be found in narrative reviews.

Conclusions—The initiative taken by NHS Research and Development in examining methods to promote the uptake of research findings is welcome, but will only prove helpful if careful attention is paid to the different stages of the development process, and different research approaches are used appropriately at different stages.

The combination of a growing number of important findings from large randomised trials and systematic overviews in areas such as the treatment of heart failure, antithrombotic therapy, and thrombolytic therapy and the realisation that the messages from research do not translate automatically into practice has led to increased interest in the means through which research results may lead to benefits for patients. In fact, an awareness of the “problem” of research implementation is not new. As David Eddy stated, the medical profession has placed high value on developing the basic science of medicine; it has not emphasised the process by which that science is translated into practice.

The English NHS has recognised explicitly the need to focus efforts on biomedical research, health services research, and implementation and has recently started a programme of research focusing largely on randomised and quasi experimental studies of interventions to change practice. The purpose of this paper is to review critically the contribution that this general approach may make to the development and evaluation of methods intended to enable patients to benefit from the important findings of biomedical and health services research. Our main focus is upon the evaluation of the effectiveness of interventions in achieving their intended purpose. However, we first discuss, and note the importance of, the qualitative work that must be undertaken to develop appropriate interventions, work that will be undertaken before experimental or quasi experimental studies are likely to be helpful. Secondly, we review the characteristics of experimental and quasi experimental studies and the interpretation of their results. Thirdly, we examine the potential role of systematic reviews and meta analysis of experimental and quasi experimental studies of interventions that aim to help health professionals provide more effective and efficient care.

Asking the right questions

DOING THE DEVELOPMENTAL WORK FIRST

A fundamental issue in conducting research in any field is asking the right questions and then choosing the right methods to answer them. This statement, although blindingly obvious,
has its basis in the importance of selecting appropriate hypotheses to test in evaluative research—hypotheses that probably will not derive originally from formal experimental research, but from practice, good ideas and their exploration in rigorous qualitative research. With the benefit of hindsight it is possible to identify areas where quantitative research has been conducted that might have been considered unnecessary or poorly focused had qualitative work been used to judge whether the hypotheses under test were sensible. For example, the perceived need for information by primary care physicians was identified in one small trial measuring intermediate outcome measures, and then its provision subsequently evaluated within a larger randomised trial with objective measurement of patient and health professional outcomes.

The outcome of the trial was that the simple provision of information to doctors did not, unfortunately, seem to bring substantial direct benefits to patients, a conclusion confirmed in a subsequent systematic review of this question. However, in a participant observational study, Covell et al. identified that the perceived need for better information among office based practitioners was not reflected in their behaviour. Although doctors perceived that they answered questions that arose from their work through traditional information sources such as text books and journal articles, they actually resolved such issues through consultation with colleagues. Thus, it was a priori unlikely that the provision of information by itself would be effective. In other words, the quantitative evaluation was ill focused when considered in the light of the relevant qualitative research in this case.

**PARALLELS WITH PHARMACEUTICAL DEVELOPMENT**

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

In pharmaceutical development this entire process is accompanied by a whole range of studies of different sorts. These vary according to the stage that has been reached, and the extent and strength of the prior beliefs about the possible outcomes. For instance, at an early stage, work may be done in a purely exploratory spirit, to generate the experience on which tentative theories can be built. Here, there may be quite well formed prior expectations with the experimenter looking for evidence of flaws in existing theories. On the other hand, she or he might simply be casting the net wide in an area where little is known. At a later stage, when a lead compound has been found, experiments will be done with the aims of increasing understanding of its characteristics and modes of action, and also of optimising them. These experiments serve both to gener-

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

...ate new avenues for exploration, and to test the many provisional hypotheses that exist at this point. By the time phase 3 clinical trials are underway, a considerable body of knowledge will have been assembled and there should be a very strong belief that the drug will work in practice. The experiments now have the focused aim of estimating the efficacy of the drug under test, and its side effects, with sufficient accuracy to support a product licence application. The varying aims of all these experiments lead to considerable differences in their design and conduct.

EXPLANATORY OR PRAGMATICALLY DESIGNED EVALUATIONS

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

A simplified model of the development process, applicable both to the clinical and behaviour change fields, consists of identifying both the underlying treatment effect and its usefulness in real world settings. This implies a corresponding classification of experiments into just two groups according to their aims. These have been called “explanatory” and “pragmatic” experiments, with the first having the aim of increasing understanding of the treatments under study, while the second are intended to estimate the size of their effects in the setting in which they would be used in practice. The principal utility of this classification is that these different aims have some general implications for design and analysis. For instance, in an “explanatory trial”, much care may be taken in controlling or measuring extraneous factors that might influence the treatment effect, while in a “pragmatic” trial, close control is neither possible or desirable.

The distinction can also be used to help crystallise the purpose of an evaluative trial beyond the ill specified wish to “test out” the treatments of interest, as the two types of experiment have conflicting strengths. For instance, explanatory studies have high construct validity (you study what is going on and aim to identify causally how something works). The disadvantage of explanatory trials is that they can often achieve high construct validity through introducing practices that are a long way from standard practice. An example of this would be the SOLVD trial that examines the effect of angiotensin converting enzyme inhibitors in patients with low ejection fractions established largely through echocardiography. In practice, many clinicians do not have direct access to echocardiography, but still have to decide whether or not to treat patients that they suspect may be similar to those in the SOLVD trial. On the other hand, pragmatic trials, in attempting to evaluate the efficiency of interventions in real world settings, strive to avoid such artificialities. However, they do this frequently at the cost of construct validity (understanding of causation) and generalisability (the extent to which the findings would apply to other treatment settings is unclear).

In summary, identifying the aims of an experiment is vital in determining a sensible protocol. Usually, questions of different types cannot be tackled efficiently in a single experiment and it is a mistake to try. Thus, what is generally needed is a structured development process, assisted by a sequence of different sorts of study or experiment. The “explanatory/pragmatic” model of experiments is often useful as a guide to thinking about these issues. However, it should be remembered that this classification is a broad one and, like all models, is not the whole story. It does not cover all types of experiment and additionally the boundary between “explanatory” and “pragmatic” trials is not rigid. In the last analysis, there is no substitute for thinking about each new case as it arises, carefully defining the aims, and being guided by these and the practical constraints that apply.

Experimental and quasi experimental studies

There are two main forms of experimental designs that may provide valid estimates of the effect of interventions—randomised trials and non-equivalent group designs. In randomised trials, subjects are allocated by chance to either treatment or control groups, and so bias is distributed between the groups by chance. In other words, in randomised trials treatment and control groups may be considered equivalent apart from the play of chance. Non-equivalent group designs are those studies in which subjects are allocated using alternatives to random allocation, and thus differences between the groups may reflect systematic bias.

ADVANTAGES OF RANDOMISATION

Randomised trials are rightly regarded as the most valid means to provide a quantitative estimate of the effectiveness of health care interventions, and randomisation plays a central part in ensuring the validity of such trials. Often, the experimenter can identify factors, other than the treatments of interest, which will affect the outcome of an experiment. Sometimes these can be controlled, or eliminated from the estimate of treatment effect by good design, thus leading to an increase in precision. However, there comes a point beyond which this is not worth the trouble, and there will also always be unknown factors at work. As such factors cannot be eliminated or designed out of the study, a concern is that they might bias the estimate of treatment effect by influencing the outcome for one group more than another.
Here, randomisation ensures that any extraneous factors have an equal chance of affecting each treatment group, which is really the best that can be done. It does not ensure that each treatment group will be exactly equally affected—as allocation depends upon the play of chance. Thus results may also be expected to differ according to chance. However, as the process of randomisation also ensures that we have a valid estimate of the expected size of this variation (through the estimate of experimental error), randomisation can be said to have ensured the validity of the statistical analysis as well. Thus, in situations where randomisation is practical, there seems no good reason for omitting it. Unfortunately, after the randomisation has been carried out, the groups have an opportunity to diverge simply as a consequence of being labelled differently, rather than as a result of a specific treatment effect. In drug trials, this problem is resolved by making them double blind, which, if done successfully, should be a complete answer. However, for professional behaviour change trials, blindness is not an option, and so this possibility needs to be tackled in a different way.

REPLICATION IN TRIALS
Another important question of design is the amount of replication to be used in an experiment. High replication ensures high precision, which, if coupled with randomisation to ensure unbiasedness, results in high accuracy. Naturally, careful consideration should be given to the degree of accuracy appropriate to any given study. The key word is appropriate—experiments can be too big as well as too small. However, for very small experiments, an additional problem arises. If replication is very low, there may be insufficient degrees of freedom on which to base the estimate of error. Here, the usual statistical analysis becomes too unstable to make much sense, and the experiment is better regarded as a non-equivalent group design, unless it can be pooled with a collection of similar studies in a meta analysis.

The requirement for valid and precise estimates of small treatment effects has led to the development of extremely large clinical trials, an example being ISIS-4 where more than 50,000 patients were randomised.22 Few trials of interventions that aim to implement research findings randomise large numbers of subjects, although effect sizes for many interventions may be expected to be moderate at best. For example, Lomas et al randomised only 16 hospitals to active interventions or control in their influential investigation of the impact on practice of “opinion leaders” or audit and feedback.23 Indeed, replication in such trials can often be lower still, for instance Norton and Dempsey randomised only six doctors to two different educational interventions.24

ESTIMATION AND CONFIDENCE INTERVALS
It is usually recommended that sample size be determined through a “power calculation”, which returns the chance of identifying real effects of a given size as being “statistically significant” at some particular level. However, this can often seem a somewhat artificial process, and particularly so in professional behaviour change trials. Here, the usual “null hypothesis” of no treatment effect is neither interesting nor particularly plausible—you are generally not evaluating methods that are thought likely to have no effect at all. Thus, a significance test is not appropriate for summarising the outcome. The question of interest is more often whether a particular intervention is worth introducing in practice or worthy of further study. Thus the concern is with the size of treatment effects. This implies that summarisation of such experiments using confidence intervals, and sample size calculations directed at predicting the probable width of such intervals, are more appropriate than significance tests and power calculations. With this approach, the experiment is viewed more like a measuring instrument that may or may not be precise enough, rather than as a diagnostic test that delivers a dichotomous answer (that is, “significant” or “non-significant”).

UNIT OF ANALYSIS
Some studies that randomise health professionals, either individually or in groups, achieve greater apparent precision through analysing patients or episodes of care. For example, Stross and Bole\textsuperscript{25} randomised six communities to receive an intervention or not, but analysed at the level of patients. While there may be situations where it is appropriate to allocate subjects within blocks, and then analyse them separately (perhaps allocating hospitals and analysing doctors) this question is outside the scope of this paper. However, it is worth pointing out that such an analysis requires additional assumptions while the standard analysis (where the units of allocation and analysis are the same) is justified by the simple act of randomisation itself. It is, in any case, an exercise in self deception to allocate doctors to behavioural interventions but analyse their patients as if they were individual unrelated events.\textsuperscript{26} This is clear almost by definition—if there is no consistency in the behaviour of a doctor from one patient to the next to start with, there seems little point in trying to change it in a consistent fashion. There is currently no adequate rationale for appropriate units for allocation and analysis in studies that attempt to influence professional practice, and given the interest and activity in this area, such guidance is overdue.

CONFOUNDING FACTORS IN NON-EQUIVALENT GROUP DESIGN STUDIES
Although potentially useful, non-equivalent group designs require more careful interpretation than large randomised studies. There are three main potential biases that may affect such studies. Firstly, there may be scaling differences in which apparent differences in effect sizes between groups over time are in fact associated with the instrument of measurement rather than real differences in effect. Secondly, it may be that different groups are at different stages in a common maturational process, and thus
respond to interventions in ways that may exaggerate or mask the true treatment effect. In other words one group may be merely lagging behind another in time, and an effect—either beneficial or otherwise—apparently associated with an intervention, may be spurious. Thirdly, the likelihood of chance differences may be high, and groups may at any time be at different positions on a common (often cyclical) process and, because outliers in such groups may be expected over time to regress to the mean, apparent differences may simply reflect this underlying chance process. These factors make the attribution of apparent responses to treatments or interventions problematic.21

COLLECTION OF ADDITIONAL DATA POINTS
One partial solution to the problem of attribution caused by non-equivalence is to collect additional data points at time periods before and after the intervention.22 A study in which there are only post intervention data points is potentially very misleading, as any contribution resulting from the intervention cannot be disentangled from what might have happened anyway. Adding extra data points before and after the intervention enables some estimation of concurrent trends and the contribution, if any, of the intervention.

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

CONTEXTUAL FACTORS
Another potential challenge to the validity of non-randomised, quasi experimental studies is the extent to which “treatments” may be separated from their context. The underlying principle of experimentation is that the manipulation of “treatments” is used to examine their
effects.21 In other words, the treatment is induced by the experimenter and the impact, or otherwise, is assessed. Where treatments are allocated randomly there is a strong likelihood that the treatment may be separated from its context. Where treatments are allocated in different ways (non-randomly) this may not be the case. Even in randomised trials, selection bias for some time been acknowledged as a potentially important bias for both those receiving and providing treatment.23 Just as practitioners may be reluctant to treat patients “at the toss of a coin”, they may also be reluctant to receive interventions intended to help them provide more effective and efficient care on this basis.

In the example of surgery for benign prostatic hyperplasia above (fig 1), the introduction of video disks may be a reflection of existing concerns on the rate of intervention—a symptom rather than a cure for the situation. In a further example in which the impact of a mass media campaign upon rates of hysterectomy in the Ticino Canton in Switzerland were examined,24 the campaign of information on the inappropriateness of high levels of intervention seems to have arisen spontaneously from concerns on rates of intervention through a form of media advocacy25 rather than having been induced by the investigators. As the informational campaign arose spontaneously rather than being induced, the comparator, Canton Bern, was clearly not equivalent at least in this important feature. Similarly, although the annual rate of hysterectomy decreased after the mass media campaign in Ticino, the rate of decrease was actually lower after the start of the campaign than in the year immediately preceding it (see fig 2). The investigators statement, that “there seems to be no doubt that the decrease in hysterectomies in Ticino was the result of the mass media information campaign”, seems heroic, not least as there was only a single data point after intervention in the neighbouring control canton, which had a much higher rate of intervention in any case.

Systematic reviews
Given the range and number of caveats raised in the previous section on interpreting the sort of evidence that may be available on the effectiveness of interventions intended to translate available evidence on clinical and cost effec-
tiveness into changes in practice, it is clear that such evidence should be interpreted with some caution. However, it is equally clear that policy decisions on the introduction of programmes intended to influence practice could also most usefully be influenced by a careful interpretation of all available evidence, rather than a subset. Missing data potentially leads to substantial bias in systematic reviews, and the best defence against this bias is a thorough and sensitive search strategy and not succumbing to the temptation to exclude studies with extreme results.

As we have discussed, there are real concerns about the validity of estimates of effect from non-randomised studies. Similarly, studies that randomise very small numbers of subjects may produce biased estimates of effect. Therefore, a reviewer may opt to include only larger randomised studies in his or her review although this strategy may produce reviews with very few studies. This approach has been taken with some effect in other contexts where there was the luxury of a comparatively large number of large studies. An alternative strategy might be to include smaller randomised and non-randomised studies and examine the robustness of the conclusions drawn from the review by sensitivity analysis on study design. Quasi-experimental studies may yield a reasonably accurate effect size in comparison with randomised designs. Further empirical work to compare the validity of treatment effects detected by different designs in a broader range of research topics would be welcome.

Reviews that include pragmatically designed studies may anticipate a priori that treatment effects can differ between studies because of interactions with other potential influences on practice. For example, there may be no a priori reason to assume that educational outreach aiming to influence prescribing decisions will act similarly among doctors in a wide range of different health care systems where both doctors and patients may face very different incentives that affect their use of drugs. In such circumstances, it may be better to use random rather than fixed effects models to combine data from various studies, as the random effects model incorporates this heterogeneity, however small, in the analysis of the overall effect of the treatment. However, as smaller trials with extreme effect sizes may exert considerable influence in such models, the sensitivity of estimates of overall effect to choice of model or inclusion of individual studies may usefully be examined. The approach often advocated in meta-analyses of deciding which model to use on the basis of a significance test for heterogeneity is compromised if the power of the statistical test is low. That approach may be more appropriate in reviews in clinical areas facing the expectation of homogeneity of effects, rather than in behavioural areas where this expectation may not be present.

The temptation to over-interpret the findings is real and is based on a lack of understanding of the limitations of the evidence. The temptation to over-interpret pragmatically designed trials should be avoided. Pragmatic studies can identify whether interventions work in the real world, but this is frequently at the cost of construct validity. Most evaluations identified in existing reviews are pragmatically designed. Pragmatically designed studies undertaken in a range of different settings do provide some information on context. However, interpretation of the apparent effects of context in these studies is a form of sub group analysis and should be reserved either for simple reassurance that an effect appears still to be present in different settings or to generate hypotheses for future research.

Conclusions
There has been comparatively little attention to the means in which research may translate to changes in practice, and this has been recognised in the programme of research from the Department of Health in England. Randomised trials, and quasi experimental designs, may provide valid quantitative estimates of the effects of interventions intended to promote the uptake of evidence based practice, but they should only be undertaken where appropriate development work using qualitative methods has been done. It is unlikely that single experiments will provide definitive answers to questions of interest, and systematic reviews of implementation research using quantitative methods to supplement narrative or descriptive reviews should be undertaken in potentially important areas. The current Department of Health programme aiming to develop means to implement research findings is only likely to prove a useful investment and lead to benefits for patients if research methodologies...
are chosen appropriately at different stages in the development of interventions, and undue reliance is not put on any particular study design. Once the qualitative work has been done, experimental designs may contribute substantially to our understanding of the effectiveness of interventions intended to influence practice and so assist health services to achieve their aims.

Funding: NHS Standing Group on Health Technology Assessment.

Conflicts of interest: none.


