Why we need observational studies to evaluate the effectiveness of health care

Nick Black

The view is widely held that experimental methods (randomised controlled trials) are the "gold standard" for evaluation and that observational methods (cohort and case control studies) have little or no value. This ignores the limitations of randomised trials, which may prove unnecessary, inappropriate, impossible, or inadequate. Many of the problems of conducting randomised trials could often, in theory, be overcome, but the practical implications for researchers and funding bodies mean that this is often not possible. The false conflict between those who advocate randomised trials in all situations and those who believe observational data provide sufficient evidence needs to be replaced with mutual recognition of the complementary roles of the two approaches. Researchers should be united in their quest for scientific rigour in evaluation, regardless of the method used.

Despite the essential role of observational methods in shedding light on the effectiveness of many aspects of health care, some scientists believe such methods have little or even nothing to contribute. In his summing up at a major conference held in 1993, the eminent medical epidemiologist Richard Doll concluded that observational methods "provide no useful means of assessing the value of a therapy."[1] The widely held view that experimental methods (randomised controlled trials) are the "gold standard" for evaluation has led to the denigration of non-experimental methods, to the extent that research funding bodies and journal editors automatically reject them. I suggest that such attitudes limit our potential to evaluate health care and hence to improve the scientific basis of how to treat individuals and how to organise services.

My main contention is that those who are opposed to the use of observational methods have assumed that they represent an alternative to experimentation rather than a set of complementary approaches. This in turn stems from a misguided notion that everything can be investigated using a randomised controlled trial. In response I want to outline the limitations of randomised trials and show that observational methods are needed both to evaluate the parts randomised trials cannot reach and to help design and interpret correctly the results obtained from these trials.

Before doing so I must clarify what I mean by "observational" in this context. I am referring exclusively to quantitative, epidemiological methods and not qualitative, sociological methods in which data are collected through observation. The principal observational epidemiological methods are non-randomised trials, cohort studies (prospective and retrospective), and case-control methods, though relatively little use has been made of the latter beyond evaluating preventive measures.

The limitations of randomised trials can be seen as deriving from either the inherent nature of the method (a limitation in principle) or from the way trials are conducted (a limitation in procedure). The importance of this distinction is that while little can be done about the former, improvements in the conduct of randomised trials could, in theory, overcome some or all of the latter. I will return to this distinction later, but first it is necessary to document the reasons why observational methods are needed. There are four main reasons: experimentation may be unnecessary, inappropriate, impossible, or inadequate.

Experimentation may be unnecessary

When the effect of an intervention is dramatic, the likelihood of unknown confounding factors being important is so small that they can be ignored. There are many well known examples of such interventions: penicillin for bacterial infections; smallpox vaccination; thyroxine in hypothyroidism; vitamin B12 replacement; insulin in insulin dependent diabetes; anaesthesia for surgical operations; immobilisation of fractured bones. In all these examples observational studies were adequate to demonstrate effectiveness.

Experimentation may be inappropriate

There are four situations in which randomised trials may be inappropriate. The first is that they are rarely large enough to measure accurately infrequent adverse outcomes. This limitation has been addressed by the establishment, in many countries, of postmarketing surveillance schemes to detect rare adverse effects of drugs. The use of such observational data can be illustrated by the case of benoxaprofen (Opren), a drug launched in 1980. Despite preceding clinical trials on over 3000 patients, the drug had to be withdrawn two years after its launch because of reports of serious side effects, including 61 deaths.[2] Similar surveillance schemes are needed for non-pharmaceutical interventions. The lack of such schemes means that there is still uncertainty as to whether or not laparoscopic techniques are associated with an increased risk of injuries, such as bile duct damage during cholecystectomy.[3] Huge observational datasets are the only practical means of acquiring such vital information.

A second limitation, also arising from study size, is the difficulty of evaluating interventions designed to prevent rare events. Examples include accident prevention schemes and placing infants supine or on their side to sleep to prevent sudden infant death syndrome. A randomised trial would have needed a few hundred thousand babies.

A third limitation of trials is when the outcomes of interest are far in the future. Three well known examples are the long term consequences of oral contraceptives, which may not be manifest for decades; the use of hormone replacement therapy to prevent femoral fractures; and the loosening of artificial hip joints, for which a 10 to 15 year follow up is needed. The practical difficulties in maintaining such prolonged prospective studies (whether experimental or observational) are considerable, as are their costs. With luck, there will occasionally be times when a randomised trial addressing the question of current interest has already been established decades before and patients from it can then be followed up. Unfortunately such serendipity is all too rare. As a practical alternative to doing nothing, retrospective observational studies

Health Services Research Unit, Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT
Nick Black, *professor of health services research*

BMJ 1996;312:1215-8

MIKE WYNDHAM PICTURE COLLECTION

What works well in pharmacological research may not work in the messier world of clinical care

can be used to obtain some information on long term outcomes.[4]

SELF DEFEATING

Finally, a randomised trial may be inappropriate because the very act of random allocation may reduce the effectiveness of the intervention. This arises when the effectiveness of the intervention depends on the subject's active participation, which, in turn, depends on the subject's beliefs and preferences. As a consequence, the lack of any subsequent difference in outcome between comparison groups may underestimate the benefits of the intervention. For example, it is well recognised that clinical audit is successful in improving the quality of health care only if the clinicians participating have a sense of ownership of the process.[5] Such a "bottom up" approach is in stark contrast to experimentation, in which the investigator seeks to impose as much control on the subjects in the study as possible—that is, a "top down" approach. As a consequence, randomised trials of audit might find less benefit than observational studies. The same may be true for many interventions for which clinicians, or patients, or both, have a preference (despite agreeing to random allocation), and where patients need to participate in the intervention—psychotherapy, for example.[6] Many interventions to promote health or prevent disease fall into this category, particularly those based on community development. It is at least as plausible to assume that experimentation reduces the effectiveness of such interventions as to assume, as most researchers have done, that the results of observational studies are wrong.

Experimentation may be impossible

There are some people who believe that any and every intervention can be subjected to a randomised trial, and that those who challenge this have simply not made sufficient effort and are methodologically incompetent. Such a view minimises the impact of seven serious obstacles that researchers have to face all too often. The exact nature of the obstacles will depend on the cultural, political, and social characteristics of the situation and, clearly, therefore, will vary over time.

The first, and most familiar, is the reluctance and refusal of clinicians and other key people to participate. Just because clinical uncertainty, manifest by variation in practice, may exist, this does not mean that each individual clinician is uncertain about how to practise. In 1991 most gynaecologists and urologists in the North Thames region agreed that a randomised trial was needed to investigate the effectiveness of surgery for stress incontinence, but none was prepared to participate as each believed in the correctness of their own practice style. In other words, although "collective equipoise" existed, "individual equipoise" was absent.[7] Even when clinicians purport to participate, randomisation may be subverted by clinicians deciphering the assignment sequence.[8]

Ethical objections are a second potential obstacle. It is most unlikely that any ethics committee in an industrialised country would sanction the random allocation of patients to intensive care versus ward care, or cardiac transplantation versus medical management. Observational studies provide an alternative to leaving the question of the effectiveness of these expensive services unevaluated. Furthermore, the results of such studies may generate sufficient uncertainty as to make an experimental study acceptable. This happened in the case of surgery for benign prostatic hyperplasia.[9 10]

POLITICAL AND LEGAL OBSTACLES

Thirdly, there may be political obstacles if those who fund and manage health services do not want their policies studied. In the United Kingdom this was true for general practitioner fundholding and the introduction of an internal market. As a result, researchers have been able to perform only a few observational studies, mostly with retrospective controls.[11 12]

Researchers may also meet legal obstacles to performing a randomised trial. The classic example is the attempt to subject radial keratotomy (an operation to correct short sightedness) to a randomised trial in the United States.[13] The researchers were blocked by private sector ophthalmologists who faced a major loss of income if the procedure was declared "experimental" because this would have meant that health insurance companies would no longer reimburse them. As a result of legal action, the academic ophthalmologists were forced to declare the operation safe and effective and abandon any attempt at evaluation.

Fortunately, legal obstacles are rare, but a common problem is that some interventions simply cannot be allocated on a random basis. These tend to be questions of how best to organise and deliver an intervention. For example, a current consensus is that clinicians and hospitals treating a high volume of patients achieve better results than those treating a low volume.[14] If true, the policy implications for the way health services are organised are immense. While experimental methods could, in theory, help resolve this, randomisation is unlikely to be acceptable to patients, clinicians, or managers. The spectre of transporting patients to more distant facilities on a randomly allocated basis would find little support from any of the interested parties. Careful observational methods provide a means of investigating the value of regionalising services.[15]

CONTAMINATION AND SCALE

The sixth problem is that of contamination. This can take several forms. If in a trial a clinician is expected to provide care in more than one way, it is possible that each approach will influence the way they provide care to patients in the other arms of the study. Consider, for example, a randomised trial to see if explaining treatments fully to patients, rather than telling them the bare minimum, would achieve better outcomes. This would rely on clinicians being able to change character repeatedly and convincingly. Fortunately there are few Dr Jekylls in clinical practice. Randomisation of clinicians (rather than patients) may sometimes help, though contamination between colleagues may occur, and randomisation of centres

requires a much larger study at far greater cost.

The seventh and final reason why it will not always be possible to conduct randomised trials is simply the scale of the task confronting the research community. There are an immense number of health care interventions in use, added to which, most interventions have many components. Consider a simple surgical operation: this entails preoperative tests, anaesthesia, the surgical approach, wound management, postoperative nursing, and discharge practice. And these are just the principal components. It will only ever be practical to subject a limited number of items to experimental evaluation.[16] We therefore need to take advantage of other methods to try and fill in the huge gaps that are always likely to exist in the experimental published findings.

Experimentation may be inadequate

The external validity, or "generalisability," of the results of randomised trials is often low.[17] The extent to which the results of a trial are generalisable depends on the extent to which the outcome of the intervention is determined by the particular person providing the care. At one extreme the outcome of pharmaceutical treatment is, to a large extent, not affected by the characteristics of the prescribing doctor. The results of drug trials can, in the main, be generalised to other doctors and settings. In contrast, the outcome of activities such as surgery, physiotherapy, psychotherapy, and community nursing may be highly dependent on the characteristics of the provider, setting, and patients. As a consequence, unless care is taken in the design and conduct of a randomised trial, the results may not be generalisable.

There are three reasons why randomised trials in many areas of health care may have low external validity. The first is that the health care professionals who participate may be unrepresentative. They may have a particular interest in the topic or be enthusiasts and innovators. The setting may also be atypical, a teaching hospital for example. In one of the few randomised trials of surgery for glue ear undertaken in the United Kingdom, all the outpatient and surgical care was performed by a highly experienced consultant surgeon; in real life most such work is performed by relatively inexperienced junior surgeons.[18]

Secondly, the patients who participate may be atypical. All trials exclude certain categories of patients. Often the exclusion criteria are so restrictive that the patients who are eligible for inclusion represent only a small proportion of the patients being treated in normal practice. Only 4% of patients currently undergoing coronary revascularisation in the United States would have been eligible for inclusion in the trials that were conducted in the 1970s.[19] It has been suggested that the same problem will limit the usefulness of the current randomised controlled trials comparing coronary artery surgery and angioplasty.[20] Similar problems occur in trials of cancer treatment.[21] Another facet of this problem is the absence of privately funded patients from almost all randomised trials in the United Kingdom.

The problem of eligibility may be exacerbated by a poor recruitment rate. Although most trials fail to report their recruitment rate,[4] those that do suggest rates are often very low. As little is yet known about the sort of people who are prepared to have their treatment allocated on a random basis, it seems wise to assume that they may differ in important ways from those who decline to take part.

And the third and final problem in generalising the results of randomised trials is that treatment may be atypical. Patients who participate may receive better care, regardless of which arm of the trial they are in.[22]

As a result of these problems, randomised trials generally offer an indication of the efficacy of an intervention rather than its effectiveness in everyday practice. While the latter can be achieved through "pragmatic" trials which evaluate normal clinical practice, these are rarely undertaken.[23] Most randomised trials are "explanatory"—that is, they provide evidence of what can be achieved in the most favourable circumstances.

The question of external validity has received little attention from those who promote randomised trials as the gold standard. None of the 25 instruments that have been developed to judge the methodological quality of trials includes any consideration of this aspect.[24] The same is true for the guidance provided by the Cochrane Collaboration.[25]

Discussion

Randomised controlled trials occupy a special place in the pantheon of methods for assessing the effectiveness of health care interventions. When appropriate, practical, and ethical, a randomised trial design should be used. I have tried to show that, for all their well known methodological strengths, trials cannot meet all our needs as patients, practitioners, managers, and policy makers. There are situations in which the use of randomised trials is limited either because of problems that derive from their inherent nature or from practical obstacles. While nothing can be done to remedy the former, improvement in the design and execution of trials could, in theory at least, overcome the latter.

PRINCIPLES VERSUS PRACTICE

The problems that could in theory be overcome (and how that could be achieved) include:

- Failure to assess rare outcomes (by mounting large trials with thousands of patients)
- Failure to assess long term outcomes (by continuing to follow up patients for many years)
- Elimination of clinicians' and patients' preferences (by introducing preference arms[26])
- Refusal by clinicians to participate (by using more acceptable methods of randomisation[27])
- Ethical objections to randomisation (by exploring alternative less demanding methods of obtaining informed consent[28])
- Political and legal obstacles (by persuasion)
- The daunting size of the task (by vastly expanding the available funds for experimental studies)
- Overrestrictive patient eligibility criteria (by undertaking pragmatic rather than explanatory trials[23])

While all the proposed solutions could work in theory, few of them are realistic in practice, presenting as they do enormous problems for researchers and, more importantly, for research funding bodies. For example, it is feasible to randomise tens of thousands of people in a drug trial in which death is the only outcome of interest, but it is unrealistic if more complex and sophisticated outcomes are the relevant endpoints. In many ways the problems that randomised trials encounter arise from a largely uncritical transfer of a well developed scientific method in pharmacological research to the evaluation of other health technologies and to health services.

Several of the other limitations cannot be polarised between principle and practice but are a complex mix of the two. These include:

- Contamination between treatment groups
- The unrepresentativeness of clinicians who volunteer to participate

- Poor patient recruitment rates
- The better care that trial participants receive

In theory all of these could be overcome, although in practice it is hard to see how without the cost of the study becoming astronomical.

Assuming procedural problems could be overcome, two problems of principle inherent in the method would remain. Firstly, the artificiality of a randomised trial probably reduces the placebo element of any intervention. Given that the placebo effect accounts for a large proportion of the effect of many interventions, the results of a trial will inevitably reflect the minimum level of benefit that can be expected. This may be one reason (along with confounding) why experimental studies often yield smaller estimates of treatment effects than studies using observational methods.[29]

Secondly, a randomised trial provides information on the value of an intervention shorn of all context, such as patients' beliefs and wishes and clinicians' attitudes and beliefs, despite the fact that such aspects may be crucial to determining the success of the intervention.[30] In contrast, observational methods maintain the integrity of the context in which care is provided. For these two reasons, the notion that information from randomised trials represents a gold standard, while that derived from observational studies is viewed as wrong, may be too simplistic. An alternative perspective is that randomised trials provide an indication of the minimum effect of an intervention whereas observational studies offer an estimate of the maximum effect. If this is so then policymakers need data from both approaches when making decisions about health services, and neither should reign supreme.

REDRESSING THE BALANCE

My intention in focusing on the limitations of trials is not to suggest that observational methods are unproblematic but to redress the balance. The shortcomings of non-experimental approaches have been widely and frequently aired. The principal problem is that their internal validity may be undermined by previously unrecognised confounding factors which may not be evenly distributed between intervention groups. It is currently unclear how serious and how insurmountable a methodological problem this is in practice. While some investigations of this issue have been undertaken,[19] more studies comparing experimental and observational designs are urgently needed.

For too long a false conflict has been created between those who advocate randomised trials in all situations and those who believe observational data provide sufficient evidence. Neither position is helpful. There is no such thing as a perfect method; each method has its strengths and weaknesses. The two approaches should be seen as complementary. After all, experimental methods depend on observational ones to generate clinical uncertainty; generate hypotheses; identify the structures, processes, and outcomes that should be measured in a trial; and help to establish the appropriate sample size for a randomised trial. When trials cannot be conducted, well designed observational methods offer an alternative to doing nothing. They also offer the opportunity to establish high external validity, something that is difficult to achieve in randomised trials.

Instead of advocates of each approach criticising the other method, everyone should be striving for greater rigour in the execution of research, regardless of the method used.

"Every research strategy within a discipline contributes importantly relevant and complementary information to a totality of evidence upon which rational clinical decision-making and public policy can be reliably based. In this context, observational evidence has provided and will continue to make unique and important contributions to this totality of evidence upon which to support a judgment of proof beyond a reasonable doubt in the evaluation of interventions."[31]

I thank Nicholas Mays, Martin McKee, Colin Sanderson, and the reviewer for their comments, but I take full responsibility for the views expressed in this article.

Funding: None.

Conflict of interest: None.

1 Doll R. Summation of conference. Doing more good than harm: the evaluation of health care interventions. *Ann N Y Acad Sci* 1994;**703**:313.
2 Opren scandal. *Lancet* 1983;i:219-20.
3 Downs SH, Black NA, Devlin HB, Royston C, Russell C. A systematic review of the safety and effectiveness of laparoscopic cholecystectomy. *Ann R Coll Surg Engl* 1996;**78**:211-23.
4 Stauffer RN. Ten-year follow-up study of total hip replacement. *J Bone Joint Surg Am* 1982;**64**:983-90.
5 Black NA. The relationship between evaluative research and audit. *J Public Health Med* 1992;**14**:361-6.
6 Brewin CR, Bradley C. Patient preferences and randomised clinical trials. *BMJ* 1989;**299**:313-15.
7 Lilford R, Jackson J. Equipoise and the ethics of randomization. *J R Soc Med* 1995;**88**:552-9.
8 Schulz KF. Subverting randomization in controlled trials. *JAMA* 1995;**274**: 1456-8.
9 Fowler FJ, Wennberg JE, Timothy RP, Barry MJ, Mulley AG, Hanley D. Symptom status and quality of life following prostatectomy. *JAMA* 1988;**259**:3018-22.
10 Wasson JH, Reda DJ, Bruskewitz RC, Elinson J, Keller AM, Henderson WG. A comparison of transurethral surgery with watchful waiting for moderate symptoms of benign prostatic hyperplasia. *N Engl J Med* 1995;**332**:75-9.
11 Dixon J, Glennerster H. What do we know about fundholding in general practice? *BMJ* 1995;**311**:727-30.
12 Clinical Standards Advisory Group. *Access to and availability of coronary artery bypass grafting and coronary angioplasty.* London: HMSO, 1993.
13 Chalmers I. Minimizing harm and maximizing benefit during innovation in health care: controlled or uncontrolled experimentation? *Birth* 1986;**13**: 155-64.
14 Black NA, Johnston A. Volume and outcome in hospital care: evidence, explanations and implications. *Health Service Management Research* 1990;**3**: 108-14.
15 Sowden AJ, Deeks JJ, Sheldon TA. Volume and outcome in coronary artery bypass graft surgery: true association or artefact? *BMJ* 1995;**311**:115-18.
16 Dorey F, Grigoris P, Amstutz H. Making do without randomised trials. *J Bone Joint Surg Br* 1994;**76-B**:1-3.
17 Cross design synthesis: a new strategy for studying medical outcomes? *Lancet* 1992;**340**:944-6.
18 Maw R, Bawden R. Spontaneous resolution of severe chronic glue ear in children and the effect of adenoidectomy, tonsillectomy, and insertion of ventilation tubes (grommets). *BMJ* 1993;**306**:756-60.
19 Hlatky MA, Califf RM, Harrell FE, Lee KL, Mark DB, Pryor D. Comparison of predictions based on observational data with the results of randomised controlled trials of coronary artery bypass surgery. *J Am Coll Cardiol* 1988;**11**:237-45.
20 White HD. Angioplasty versus bypass surgery. *Lancet* 1995;**346**:1174-5.
21 Ward LC, Fielding JWL, Dunn JA, Kelly KA. The selection of cases for randomised trials: a registry of concurrent trial and non-trial participants. *Br J Cancer* 1992;**66**:943-50.
22 Stiller CA. Centralised treatment, entry to trials, and survival. *Br J Cancer* 1994;**70**:352-62.
23 Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in clinical trials. *J Chron Dis* 1967;**20**:637-48.
24 Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Controlled Clin Trials* 1995;**16**:62-73.
25 Oxman AD, ed. Section VI: Preparing and maintaining systematic reviews. *Cochrane Collaboration handbook.* Oxford: Cochrane Collaboration, 1994.
26 Wennberg JE, Barry MJ, Fowler FJ, Mulley A. Outcomes research, PORTs, and health care reform. Doing more good than harm: the evaluation of health care interventions. *Ann NY Acad Sci* 1994;**703**:56.
27 Korn EL, Baumrind S. Randomised clinical trials with clinician-preferred treatment. *Lancet* 1991;**337**:149-52.
28 Zelen M. A new design for randomised clinical trials. *N Engl J Med* 1979;**300**:1242-5.
29 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* 1995;**273**:408-12.
30 Beecher HK. Surgery as placebo. *JAMA* 1961;**176**:1102-7.
31 Hennekens CH, Buring JE. Observational evidence. Doing more good than harm: the evaluation of health care interventions. *Ann NY Acad Sci* 1994;**703**:22.

(Accepted 7 March 1996)