

- 7 Forman D, Rider L, eds. *Yorkshire Cancer Registry report 1996*. Leeds: Yorkshire Cancer Organisation, 1996.
- 8 Campbell AJ, Buchner DM. Unstable disability and the fluctuations of frailty. *Age Ageing* 1997;26:315-8.
- 9 Balducci L, Mowrey K, Parker M. Pharmacology of antineoplastic agents in older patients. In: Balducci L, Lyman GH, Ershler WB, eds. *Geriatric oncology*. Philadelphia: JB Lippincott Co, 1992:169-80.
- 10 Shank WA, Jr., Balducci L. Recombinant hemopoietic growth factors: comparative hemopoietic response in younger and older subjects. *J Am Geriatr Soc* 1992;40:151-4.
- 11 Begg CB, Carbone PP. Clinical trials and drug toxicity in the elderly. The experience of the Eastern Cooperative Oncology Group. *Cancer* 1983;52:1986-92.
- 12 Evans WK, Radwi A, Tomiak E, Logan DM, Martins H, Stewart DJ, et al. Oral etoposide and carboplatin. Effective therapy for elderly patients with small cell lung cancer. *Am J Clin Oncol* 1995;18:149-55.
- 13 Morandi U, Stefani A, Golinelli M, Ruggiero C, Brandi L, Chiapponi A, et al. Results of surgical resection in patients over the age of 70 years with non small-cell lung cancer. *Eur J Cardiothorac Surg* 1997;11:432-9.
- 14 Bufalari A, Ferri M, Cao P, Cirocchi R, Bisacci R, Moggi L. Surgical care in octogenarians. *Br J Surg* 1996;83:1783-7.
- 15 Farrow DC, Hunt WC, Samet JM. Temporal and regional variability in the surgical treatment of cancer among older people. *J Am Geriatr Soc* 1996;44:559-64.
- 16 Pignon T, Gregor A, Schaake KC, Roussel A, Van Glabbeke M, Scalliet P. Age has no impact on acute and late toxicity of curative thoracic radiotherapy. *Radiother Oncol* 1998;46:239-48.
- 17 Pignon T, Horiot JC, Bolla M, van Poppel H, Bartelink H, Roelofs F, et al. Age is not a limiting factor for radical radiotherapy in pelvic malignancies. *Radiother Oncol* 1997;42:107-20.
- 18 Rostom AY, Pradhan DG, White WF. Once weekly irradiation in breast cancer. *Int J Radiat Oncol Biol Phys* 1987;13:551-5.
- 19 Olmi P, Ausili-Cefaro G. Radiotherapy in the elderly: a multicentric prospective study on 2060 patients referred to 37 Italian radiation therapy centers. *Rays* 1997;22:53-6.
- 20 Robertson JF, Todd JH, Ellis IO, Elston CW, Blamey RW. Comparison of mastectomy with tamoxifen for treating elderly patients with operable breast cancer. *BMJ* 1988;297:511-4.
- 21 Balducci L, Extermann M, Fentiman I, Monfardini S, Perrone F. Should adjuvant chemotherapy be used to treat breast cancer in elderly patients (≥ 70 years of age)? *Eur J Cancer* 1997;33:1720-4.
- 22 Cleary JF, Carbone PP. Palliative medicine in the elderly. *Cancer* 1997;80:1335-47.
- 23 Fletcher A. Screening for cancer of the cervix in elderly women. *Lancet* 1990;335:97-9.
- 24 Age Concern. *Not at my age: why the present breast screening system is failing women aged 65 or over*. London: Age Concern England, 1996.
- 25 Van Dijk JA, Verbeek AL, Beex LV, Hendriks JH, Holland R, Mravunac M, et al. Mammographic screening after the age of 65 years: evidence for a reduction in breast cancer mortality. *Int J Cancer* 1996;66:727-31.
- 26 Haigney E, Morgan R, King D, Spencer B. Breast examinations in older women: questionnaire survey of attitudes of patients and doctors. *BMJ* 1997;315:1058-9.
- 27 Yellen SB, Cella DF, Leslie WT. Age and clinical decision making in oncology patients. *J Natl Cancer Inst* 1994;86:1766-70.
- 28 Mead GE, Pendleton N, Pendleton DE, Horan MA, Bent N, Rabbit P. High technology medical interventions: what do older people want? *J Am Geriatr Soc* 1997;45:1409-11.
- 29 Given CW, Given BA, Stommel M. The impact of age, treatment, and symptoms on the physical and mental health of cancer patients. A longitudinal perspective. *Cancer* 1994;74:2128-38.
- 30 Weinrich SP, Weinrich MC. Cancer knowledge among elderly individuals. *Cancer Nurs* 1986;9:301-7.
- 31 Wetle T. Age as a risk factor for inadequate treatment. *JAMA* 1987;258:516.
- 32 Newcomb PA, Carbone PP. Cancer treatment and age: patient perspectives. *J Natl Cancer Inst* 1993;85:1580-4.
- 33 Pearlman RA, Uhlmann RF. Quality of life in chronic diseases: perceptions of elderly patients. *J Gerontol* 1988;43:M25-30.
- 34 Hazzard WR, Woolard N, Regenstreif DI. Integrating geriatrics into the specialties of internal medicine: the Hartford Foundation/American Geriatrics Society/Wake Forest University Bowman Gray School of Medicine initiative. *J Am Geriatr Soc* 1997;45:638-40.
- 35 Sainsbury R, Rider L, Smith A, MacAdam A. Does it matter where you live? Treatment variation for breast cancer in Yorkshire. The Yorkshire Breast Cancer Group. *Br J Cancer* 1995;71:1275-8.
- 36 Calman K, Hine D. *A policy framework for commissioning cancer services. A report by the Expert Advisory Group on Cancer to the Chief Medical Officers of England and Wales, 1995*. London: Department of Health, 1995.
- 37 Selby P, Gillis C, Haward R. Benefits from specialised cancer care. *Lancet* 1996;348:313-8.
- 38 Monfardini S. What do we know on variables influencing clinical decision-making in elderly cancer patients? *Eur J Cancer* 1996;32A:12-4.
- 39 Fentiman IS. Are the elderly receiving appropriate treatment for cancer? *Ann Oncol* 1996;7:657-8.
- 40 Coebergh JW. Significant trends in cancer in the elderly. *Eur J Cancer* 1996;32A:569-71.

(Accepted 6 April 1999)

Methods in health services research

Interpreting the evidence: choosing between randomised and non-randomised studies

Martin McKee, Annie Britton, Nick Black, Klim McPherson, Colin Sanderson, Chris Bain

This is the first of four articles

Correspondence to: Martin McKee
m.mckee@lshtm.ac.uk

Series editor: Nick Black

continued over

BMJ 1999;319:312-5

website extra

Further references are listed on the BMJ's website

www.bmj.com

Evaluations of healthcare interventions can either randomise subjects to comparison groups, or not. In both designs there are potential threats to validity, which can be external (the extent to which they are generalisable to all potential recipients) or internal (whether differences in observed effects can be attributed to differences in the intervention). Randomisation should ensure that comparison groups of sufficient size differ only in their exposure to the intervention concerned. However, some investigators have argued that randomised controlled trials (RCTs) tend to exclude, consciously or otherwise, some types of patient to whom results will subsequently be applied. Furthermore, in unblinded trials the outcome of treatment may be influenced by practitioners' and patients' preferences for one or other intervention. Though non-randomised studies are less selective in terms of recruitment, they are subject to selection bias in allocation if treatment is related to initial prognosis.

These issues have led to extensive debate, although empirical evidence is limited. This paper is a brief summary of a more detailed review¹ of the impact of these potential threats.

Summary points

Treatment effects obtained from randomised and non-randomised studies may differ, but one method does not give a consistently greater effect than the other

Treatment effects measured in each type of study best approximate when the exclusion criteria are the same and where potential prognostic factors are well understood and controlled for in the non-randomised studies

Subjects excluded from randomised controlled trials tend to have a worse prognosis than those included, and this limits generalisability

Subjects participating in randomised controlled trials evaluating treatment of existing conditions tend to be less affluent, educated, and healthy than those who do not; the opposite is true for trials of preventive interventions

Threats to validity of evaluative research and possible solutions

Type of validity	Threatening factor	Proposed solution
Internal	Allocation bias (risk of confounding)	Randomisation
		Risk adjustment and subgroup analysis (analysis)
	Patient preference	Preference arms or adjustment for preference (design)
External	Exclusions (eligibility criteria)	Expand inclusion criteria
	Non-participation (centres/practitioners)	Multicentre, pragmatic design
	Not invited (practitioner preference or administrative oversight)	Encourage practitioners to invite all eligible patients
	Non-participation (patients)	Less rigorous consent procedures

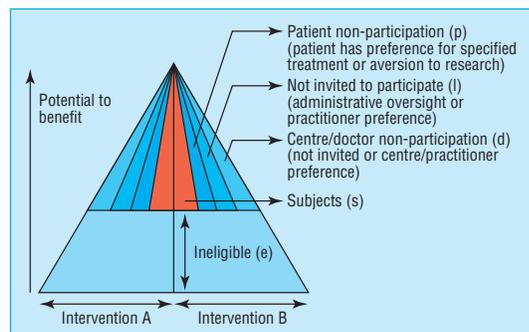


Fig 1 Differences in inclusion and participation. Shaded areas represent the study population

Nature of the evidence

The review focused on threats to internal and external validity of evaluations of effectiveness and on the strategies proposed to overcome them (table). Various factors act through their effect on the distribution of the potential to benefit among different groups. This can be illustrated schematically (fig 1). The reference population is defined by an envelope, represented here as a triangle but potentially taking many shapes. At some point, a threshold is reached, below which the overall risks outweigh the benefits. As patients are excluded or do not participate, the study population becomes a progressively smaller subset of the reference population, in principle increasing the scope for selection bias and raising the question of whether it is valid to apply the results obtained to the reference population.

We used systematic reviews to explore the potential and actual importance of factors that lead to selective recruitment, examining four questions:

- Do non-randomised studies give systematically different measurements of treatment effect from RCTs?
- Are there systematic differences between the subjects included in or excluded from studies, and do these influence the measured treatment effect?
- To what extent is it possible to overcome known or unknown baseline differences between groups that are not allocated randomly?
- How important are patients' preferences for an intervention and, if patients are randomised to a treatment they would not choose, how does this affect their outcome?

Findings

Comparing results of RCTs and non-randomised studies

Eighteen papers were identified where a single intervention was evaluated by both methods (a full list

is available on the *BMJ's* website). A review was published just after our original report; on the basis of eight comparisons it found that, on average, non-randomised studies overestimate effect size.² In contrast, of the seven studies in our review where the two methods detected effects in the same direction, in three the effect size was greater in the randomised trial and in four it was greater in the non-randomised study. The key finding in our study is that neither method consistently gave larger estimates of treatment effect.

In addition to chance, there are several potential explanations for different measurements of treatment effects. The overall impact will reflect the relative importance of each issue in a particular case. A randomised controlled trial may produce a greater effect if the patients enrolled in it receive higher quality care or are selected so that they have greater capacity to benefit than patients in non-randomised studies. But it may produce a lower estimate of treatment effect for several reasons:

- In non-randomised studies, patients tend to be allocated to treatments that are correctly considered most appropriate for their individual circumstances;
- Exclusions from a RCT create a sample with less capacity to benefit than in a non-randomised study;
- An unblinded RCT fails to capture patients with strong preferences for a particular treatment who show an enhanced response to treatment;
- Non-randomised studies of preventive interventions include disproportionate numbers of individuals who, by virtue of their health related behaviour, have greater capacity to benefit;
- Publication bias leads to negative results being less likely to be published from non-randomised studies than from RCTs.

The limited evidence indicates that the results of non-randomised studies best approximate to results of RCTs when both use the same exclusion criteria and when potential prognostic factors are well understood, measured, and appropriately controlled in non-randomised studies.³

In summary, the results of RCTs and non-randomised studies of similar patients may not, after adjustment, be substantially different in relative effect size. Any variations are often no greater than those between different RCTs or among non-randomised studies. Differences in effect sizes could be due to chance or differences in populations studied, timing, or nature of the intervention.

Exclusions

Randomised controlled trials vary widely in their inclusiveness. Medical reasons cited for exclusion from trials include a high risk of adverse effects and belief

London School of Hygiene and Tropical Medicine, London WC1E 7HT

Martin McKee
professor of European public health

Annie Britton
research fellow

Nick Black
professor of health services research

Klim McPherson
professor of public health epidemiology

Colin Sanderson
senior lecturer in health services research

University of Queensland Medical School, Brisbane 4006, Australia

Chris Bain
reader in social and preventive medicine

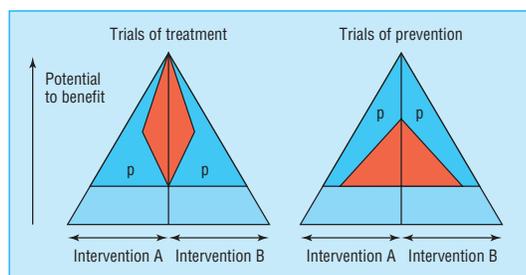


Fig 2 Effect of differences in participation in trials of prevention and of treatment. p=eligible non-participants; shaded areas represent the study population

that benefit, or lack of it, has already been established for some groups.

Scientific reasons include greater precision in estimating treatment effects by having a homogeneous sample,⁴ and reduced risk of bias by excluding individuals most likely to be lost to follow up.⁵ In addition, many RCTs have blanket exclusions,⁶ the reasons for which are often unstated, of categories of patients such as the elderly, women, and ethnic minorities.

Few studies have examined differences in prognostic factors between included and excluded patients, but some have used clinical databases to examine this.^{7,8} The patients included in such databases tended to have a poorer prognosis than those in trials: in one study, a subset selected to meet eligibility criteria of RCTs produced treatment effects of similar size to those obtained from RCTs.³

Participation

Evaluative research is undertaken predominantly in university or teaching centres, but non-randomised studies are more likely than RCTs to include non-teaching centres, and criteria for participation in RCTs may include the achievement of a specified level of clinical outcome. The available evidence suggests that this may exaggerate the measured treatment effect.⁹

Most evaluative studies fail to document adequately the characteristics of eligible patients who do not participate. The effect of non-participation differs between RCTs that evaluate interventions designed to treat an existing condition and those directed at preventing disease (fig 2).¹⁰ Participants in the former tend to be less affluent, less educated, and more severely ill than eligible patients who do not participate.¹¹ In contrast, participants in RCTs evaluating preventive interventions tend to be more affluent, better educated, and more likely to have adopted a healthy lifestyle than patients who decline.¹² On the basis of the evidence from the comparisons discussed earlier, it is plausible that low participation in RCTs of treatment may exaggerate treatment effects by including more skilful practitioners and subjects with a greater capacity to benefit, while RCTs of prevention may underestimate effects as participants have selectively less capacity to benefit.

Impact of patients' preferences

There is little empirical research on the impact on outcome of patients' preferences. The four studies that attempted to measure preference effects either were small or have yet to report full results.¹³⁻¹⁶ In theory, preference could have an important impact on results

of RCTs, especially where the true effect is small. Such effects could account for some observed differences between results of RCTs and non-randomised studies. There are methods that may detect preference effects reliably; though these may contribute to understanding this phenomenon, none provides a complete answer.¹⁷ This is mainly because randomisation between preferring a treatment and not is impossible, and confounding may bias any observed comparison.

Adjustment for baseline differences in non-randomised studies

Despite the evidence that the results of RCTs and non-randomised studies are often similar, differences in baseline prognostic factors clearly can be important. Absence of randomisation can produce groups that differ in important ways, and it is necessary to consider whether it is possible to adjust for such differences. Adjustment for imbalance in baseline prognostic factors between arms of non-randomised studies commonly changes the size of the measured treatment effect, but such changes are often small and inconsistent.¹

Overall, the limited evidence suggests that differences in the populations studied by RCTs and non-randomised studies are likely to be of at least as much importance in explaining any differences and that the two methods should be compared only after patients not meeting eligibility criteria for the RCT are excluded.

Recommendations

A large, inclusive, fully blinded RCT incorporating appropriate subgroup analysis is likely to provide the best possible evidence of effectiveness, but there will always be circumstances in which randomisation, especially on an inclusive basis, is unethical or impractical.¹⁸ In circumstances where there are genuine reasons for not randomising,¹⁹ non-randomised studies can provide useful evidence. In such studies, adjustment for baseline imbalances should always be attempted, as rigorously and extensively as possible, and the procedures should be reported explicitly to help readers' evaluations. However, adjustment cannot be relied on to approximate the prognostic balance of randomisation, given unknown or unmeasurable confounding.

Investigators conducting evaluative research (using any design) must improve the quality of reporting. Authors should define the population to whom they expect their results to be applied; what has been done to ensure that the study population is representative of this wider population, and any evidence of how it differs; whether centres that participated differ from those that declined; and the numbers and characteristics of patients eligible to be included who either were not invited to do so or were invited and declined.

The findings of such studies have implications for the way in which evidence is interpreted. When faced with data from any source, whether randomisation has been used or not, it is important first to pursue alternative (non-causal) explanations thoroughly and examine the possible influence of chance, bias, and confounding, perhaps using sensitivity analyses where feasible.

Where only non-randomised data are available, the potential for allocation bias should be considered and any attempts at risk adjustment should be assessed.

Where only randomised trials are found, preference effects should also be considered. To obtain an uncontaminated estimate of the physiological effect of a treatment, RCTs should be blind to everyone involved, but for many interventions this will be impossible. Also, the advantages of narrowing inclusion criteria to ensure high participation in RCTs should be balanced by the potential need for subgroup analysis. It should not be assumed that a summary result applies to all potential patients.

When both randomised and non-randomised studies have been conducted it is important to ascertain whether estimates of treatment effect are consistent for patients at similar risk across studies. If so, it may be reasonable to accept the results of the less exclusive non-randomised studies. Differences in results cannot be assumed to be solely due to the presence or lack of randomisation—differences in study populations, characteristics of the intervention, and the effects of patients' preferences may also affect the results.

Whichever design is used, generalisability needs attention. One approach involves examining the relation between reduction in relative risk (as a measure of effect size) against the percentage of events in the control arm (as an indirect measure of inclusiveness)²⁰; this is sometimes referred to as metaregression.²¹ Where sufficient data are available from RCTs, it may be possible to identify separate measures of benefit and harm. If, as has been shown for giving anticoagulants to prevent stroke, the percentage reduction in relative risk remains constant at all levels of severity and the increase in absolute risk of an adverse effect remains constant, the reduction in absolute risk for a given patient can then be estimated.²²

In conclusion, RCTs and non-randomised studies can provide complementary evidence—but it is important that clinicians using this evidence are aware of the strengths and weaknesses of each method.

This article is adapted from *Health Services Research Methods: A Guide to Best Practice*, edited by Nick Black, John Brazier, Ray Fitzpatrick, and Barnaby Reeves, published by BMJ Books in 1998.

Funding: This work was supported by a grant from the NHS Health Technology Assessment Programme.

Competing interests: None declared.

- 1 Britton A, McKee M, Black N, McPherson K, Sanderson C, Bain C. Choosing between randomised and non-randomised studies: a systematic review. *Health Technol Assess* 1998;2(13).
- 2 Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 1998;317:1185-90.
- 3 Horvitz RI, Viscoli CM, Clemens JD, Sadock RT. Developing improved observational methods for evaluating therapeutic effectiveness. *Am J Med* 1990; 89:630-8.
- 4 Yusuf S, Held P, Teo KK. Selection of patients for randomized controlled trials: implications of wide or narrow eligibility criteria. *Stat Med* 1990;9:73-86.
- 5 Haynes RB, Dantes R. Patient compliance and the conduct and interpretation of therapeutic trials. *Controlled Clin Trials* 1987;8:12-9.
- 6 Gurwitz JH, Nananda F, Avorn J. The exclusion of the elderly and women from clinical trials on acute myocardial infarction. *JAMA* 1992;268: 1417-22.
- 7 Jones RH, Kesler K, Phillips HR, Mark DB, Smith PK, Nelson CL, et al. Long-term survival benefits of coronary artery bypass grafting and percutaneous transluminal angioplasty in patients with coronary artery disease. *J Thorac Cardiovasc Surg* 1996;11:1013-25.
- 8 Hartz AJ, Kuhn EM, Pryor DB, Krakauer H, Young M, Heudeberg, et al. Mortality after coronary angioplasty and coronary artery bypass surgery (the National Medicare Experience). *Am J Cardiol* 1992;70: 179-85.
- 9 Stukenborg GJ. Comparison of carotid endarterectomy outcomes from randomized controlled trials and medicare administrative databases. *Arch Neurol* 1997;54:826-32.
- 10 Hunninghake DB, Darby CA, Probstfield JL. Recruitment experience in clinical trials: literature summary and annotated bibliography. *Controlled Clin Trials* 1987;8:6S-30S.
- 11 Barofsky I, Sugarbaker PH. Determinants of patients nonparticipation in randomised clinical trials for the treatment of sarcomas. *Cancer Clin Trials* 1979;2:237-49.
- 12 Davies G, Pyke S, Kinmouth AL. Effect of non-attenders on the potential of a primary care programme to reduce cardiovascular risk in the population. *BMJ* 1994;309:1553-6.
- 13 McKay JR, Alterman AI, McLellan T, Snider EC, O'Brien CP. Effect of random versus nonrandom assignment in a comparison of inpatient and day hospital rehabilitation for male alcoholics. *J Consulting Clin Psychol* 1995;63:70-8.
- 14 Nicolaides K, de Lourdes Brizot M, Patel F, Snijders R. Comparison of chorionic villus sampling and amniocentesis for fetal karyotyping at 10-13 weeks' gestation. *Lancet* 1994;344:435-9.
- 15 Torgerson DJ, Klaber-Moffett J, Russell IT. Patient preferences in randomised trials: threat or opportunity. *J Health Serv Res Policy* 1996;1:194-7.
- 16 Fallowfield LJ, Hall A, Maguire GP, Baum M. Psychological outcomes of different treatment policies in women with early breast cancer outside a clinical trial. *BMJ* 1990;301:575-80.
- 17 McPherson K, Britton AR, Wennberg JE. Are randomized controlled trials controlled? Patient preferences and unblind trials. *J Roy Soc Med* 1997;90:652-6.
- 18 Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;312:1215-8.
- 19 McPherson K. The Cochrane Lecture. The best and the enemy of the good: randomised controlled trials, uncertainty, and assessing the role of patient choice in medical decision making. *J Epidemiol Commun Health* 1994;48:6-15.
- 20 Davey Smith G, Song F, Sheldon T. Cholesterol lowering and mortality: the importance of considering initial risk. *BMJ* 1993;206:1367-73.
- 21 Holme I. Cholesterol reduction and its impact on coronary artery disease and total mortality. *Am J Cardiol* 1995;76:10-7C.
- 22 Glasziou PP, Irwig LM. An evidence based approach to individualising treatment. *BMJ* 1995;311:1356-9.



Available from the BMJ bookshop (www.bmjbookshop.com)

Email submissions from outside the United Kingdom

We are now offering an email submission service for authors from outside the UK. The address is papers@bmj.com

Ideally our email server would link seamlessly with our manuscript tracking system, but for now it does not, which is why we are offering the service only to authors outside the UK. Most post in the UK arrives the next day, so UK authors have the least to gain in speed of delivery from email delivery. As soon as our systems improve we will invite email submissions from everyone.

If you choose to send your submission by email please would you send the text and any tables and figures as attached files, together with a covering letter giving all your contact details (postal

address, phone, fax, and email address). We can read files created with most word processing, graphics, and spreadsheet programs.

When your submission is received in our email box you will receive an automatic acknowledgment to show that it has arrived. If the submission is incomplete we will contact you and ask you to resend the missing information.

Once the submission is complete we will register it on our manuscript tracking system and you will receive a standard acknowledgment in the post.

Letters to the editor should continue to be sent direct to www.bmj.com as rapid responses or to letters@bmj.com

papers@bmj.com

the codes of the active and inert drugs.³ Furthermore, treatment allocation can be guessed if blocking is used. For instance if patients are randomised in a series of blocks of four—that is, for every four patients randomised two will receive one treatment and two will receive the other—an investigator who remembers the treatments the previous three patients received will be able to predict the treatment for the fourth.

While much of the evidence on subverting randomisation is anecdotal, a recent review found that randomisation has been compromised in several controlled trials.² This review showed that trials which did not adequately conceal randomisation from the investigators demonstrated, on average, a 41% increase in effect for the active treatment compared with an adequately concealed trial.² Indeed, in a current multi-centre randomised trial of a surgical procedure in the United Kingdom the median age of patients for the experimental treatment was found to be significantly lower for three groups of clinicians when an envelope system was used. This age imbalance disappeared when better concealment measures were introduced.⁴

Owing to the problems of using envelopes it is methodologically more sound to undertake “distance”

randomisation (although in some instances sealed envelopes may be the only practical means of randomisation). Distance randomisation usually involves the investigator, on recruiting a patient, telephoning a central randomisation service which notes basic patient details and then issues a treatment allocation. Indeed, distance randomisation can now be performed over the internet. Such a system is being used, alongside telephone randomisation, in the Medical Research Council’s growth restricted intervention trial (GRIT). Distance randomisation is much less likely to be compromised than an envelope system.

Thus, to avoid bias it is important that randomisation is well concealed. Recent evidence has questioned the rigor of using local randomisation. Randomisation should be distant and separate from clinicians conducting the trial.

- 1 Pocock SJ. *Clinical trials: a practical approach*. Chichester: John Wiley, 1983.
- 2 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of effects in controlled trials. *JAMA* 1995;273:408-12.
- 3 Schulz KF. Subverting randomisation in controlled trials. *JAMA* 1995;274:1456-8.
- 4 Kennedy A, Grant A. Subversion of allocation in a randomised controlled trial. *Control Clin Trials* 1997;18(suppl 3):77-85.

Methods in health service research

Evaluation of health interventions at area and organisation level

Obioha C Ukoumunne, Martin C Gulliford, Susan Chinn, Jonathan A C Sterne, Peter G J Burney, Allan Donner

This is the second of four articles

Department of Public Health Sciences, Guy’s, King’s, and St Thomas’s School of Medicine, King’s College, London SE1 3QD

Obioha C

Ukoumunne,
research associate

Martin C Gulliford,
senior lecturer

Susan Chinn,
reader

Jonathan A C Sterne,
senior lecturer

Peter G J Burney,
professor

continued over

BMJ 1999;319:376-9

Healthcare interventions are often implemented at the level of the organisation or geographical area rather than at the level of the individual patient or healthy subject. For example, screening programmes are delivered to residents of a particular area; health promotion interventions might be delivered to towns or schools; general practitioners deliver services to general practice populations; hospital specialists deliver health care to clinic populations. Interventions at area or organisation level are delivered to clusters of individuals.

The evaluation of interventions based in an area or organisation may require the allocation of clusters of individuals to different intervention groups (see box 1).^{1,2} Cluster based evaluations present special problems both in design and analysis.³ Often only a small number of organisational units of large size are available for study, and the investigator needs to consider the most effective way of designing a study with this constraint. Outcomes may be evaluated either at cluster level or at individual level (table).⁴ Often cluster level interventions are aimed at modifying the outcomes of the individuals within clusters, and it will then be important to recognise that outcomes for individuals within the same organisation may tend to be more similar than for individuals in different organisational clusters (see box 2). This dependence between individuals in the same cluster has important implications for the design and analysis of organisation based studies.² This paper addresses these issues.

Summary points

Health interventions are often implemented at the levels of health service organisational unit or of geographical or administrative area

The unit of intervention is then a cluster of individual patients or healthy subjects

Evaluation of cluster level interventions may be difficult because only a few units of large size may be available for study, evaluation may be at either individual or cluster level, and individuals’ responses may be correlated within clusters

At the design stage, it is important to randomise clusters whenever possible, adapt sample size calculations to allow for clustering of responses, and choose between cohort and repeated cross sectional designs

Methods chosen for analysis of individual data should take into account the correlation of individual responses within clusters

Nature of the evidence

We retrieved relevant literature using computer searches of the Medline, BIDS (Bath Information and

Comparison of levels of intervention and levels of evaluation (adapted from McKinlay⁴)

Level of evaluation	Level of intervention	
	Individual	Area or organisation
Individual	Clinical trial—for example, does treating multiple sclerosis patients with interferon beta reduce their morbidity from the condition?	Area or organisation based evaluation—for example, does providing GPs with guidelines on diabetes management improve blood glucose control in their patients? Does providing a "baby friendly" environment in hospital increase mothers' success at breast feeding?
Area or organisation		Area or organisation based evaluation—for example, do smoking control policies increase the proportion of smoke free workplaces? Do fundholding general practices develop better practice facilities than non-fundholders?

Department of
Epidemiology and
Biostatistics,
University of
Western Ontario,
London, Ontario,
Canada N6A 5C1
Allan Donner,
chairman

Correspondence to:
MC Gulliford
martin.gulliford@
kcl.ac.uk

Series editor: Nick
Black

Data Services), and ERIC (Education Resources Information Centre) databases and hand searches of relevant journals. The papers retrieved included theoretical statistical studies and studies that applied these methods. Much of the relevant work has been done on community intervention studies in coronary heart disease prevention. We retrieved the content of the papers, made qualitative judgments about the validity of different approaches, and synthesised the best evidence into methodological recommendations.

Findings

We identified 10 key considerations for evaluating organisation level interventions.

(1) Recognise the cluster as the unit of intervention or allocation

Healthcare evaluations often fail to recognise, or use correctly, the different levels of intervention which may be used for allocation and analysis.⁵ Failure to distinguish individual level from cluster level intervention or analysis can result in studies that are inappropriately designed or give incorrect results.³

(2) Justify the use of the cluster as the unit of intervention or allocation

For a fixed number of participants, studies in which clusters are randomised to groups are not as powerful as traditional clinical trials in which individual patients

are randomised.² The decision to allocate at organisation level should therefore be justified on theoretical, practical, or economic grounds (box 1).

(3) Include enough clusters

Evaluation of an intervention that is implemented in a single cluster will not usually give generalisable results. For example, a study evaluating a new way of organising care at one diabetic clinic would be an audit study which may not be generalisable. It would be better to compare control and intervention clinics, but studies with only one clinic per group would be of little value, since the effect of intervention is completely confounded with other differences between the two clinics. Studies with only a few (fewer than four) clusters per group should generally be avoided as the sample size will be too small to allow a valid statistical analysis with appreciable chance of detecting an intervention effect. Studies with as few as six clusters per group have been used to show effects from cluster based interventions,⁶ but larger numbers of clusters will often be needed, particularly when relevant intervention effects are small.

(4) Randomise clusters wherever possible

Random allocation has not been used as often as it should in the evaluation of interventions at the level of area or organisation. Randomisation should be used to avoid bias in the estimate of intervention effect as a result of confounding with known or unknown factors. Sometimes the investigator will not be able to control the assignment of clusters—for instance, when evaluating an existing service,⁷ but because of the risk of bias, randomised designs should always be preferred. If randomisation is not feasible, then the chosen study design should allow for potential sources of bias.⁸ Non-randomised studies should include intervention and control groups with observations made before and after the intervention. If only a single group can be studied, observations should be made on several occasions both before and after the intervention.⁸

(5) Allow for clustering when estimating the required sample size

When observations made at the individual level are used to evaluate interventions at the cluster level, standard formulas for sample size will not be appropriate for obtaining the total number of participants required. This is because they assume that the responses of individuals within clusters are independent (box 2).^{2 9-11} Standard sample size formulas underestimate the number of participants required because they allow for variation within clusters but not between clusters.

Box 1: Reasons for carrying out evaluations at cluster level

- Public health and healthcare programmes are generally implemented at organisation rather than individual level, so cluster level studies are more appropriate for assessing the effectiveness of such programmes
- It may not be appropriate, or possible in practice, to randomise individuals to intervention groups since all individuals within a general practice or clinic may be treated in the same way
- "Contamination" may sometimes be minimised through allocation of appropriate organisational clusters to intervention and control groups. For example, individuals in an intervention group might communicate a health promotion message to control individuals in the same cluster. This might be minimised by randomising whole towns to different interventions
- Studies in which entire clusters are allocated to groups may sometimes be more cost effective than individual level allocation, if locating and randomising individuals is relatively costly

Box 2: Three reasons for correlation of individual responses within area or organisational clusters

- Healthy subjects or patients may have chosen the social unit to which they belong. For example, individuals may select their general practitioners on the basis of characteristics such as age, sex, or ethnic group. Individuals who choose the same social or organisational unit might be expected to have something in common
- Cluster level attributes may have a common influence over all individuals in that cluster, thus making them more similar. For example, outcomes of surgery may vary systematically between surgeons, so that outcomes for patients treated by one surgeon tend to be more similar to each other than to those of another surgeon
- Individuals may interact within the cluster, leading to similarities between individuals for some health related outcomes. This might occur, for example, when individuals within a community respond to health promotion messages communicated through news media

To allow for the correlation between subjects, the required standard sample size derived from formulas for individually randomised trials should be multiplied by a quantity known as the design effect or variance inflation factor.^{2,9} This will give a cluster level evaluation with the same power to detect a given intervention effect as a study with individual allocation. The design effect is estimated as

$$Deff=1+(n_0-1)\rho$$

where *Deff* is the design effect, n_0 is the average number of individuals per cluster and ρ is the intraclass correlation coefficient for the outcome of interest.

The intraclass correlation coefficient is the proportion of the total variation in the outcome that is between clusters; this measures the degree of similarity or correlation between subjects within the same cluster. The larger the intraclass correlation coefficient—that is, the more the tendency for subjects within a cluster to be similar—the greater the size of the design effect and the larger the additional number of subjects required in an organisation based evaluation, compared with an individual based evaluation.

Sample size calculations require the intraclass correlation coefficient to be known or estimated before the study is carried out.¹² If the intraclass coefficient is not available, plausible values must be estimated. A range of components of variance and intraclass correlations is reported elsewhere.^{13,14}

The number of clusters required for a study can be estimated by dividing the total number of individuals required by the average cluster size. When sampling of individuals within clusters is feasible, the power of the study may be increased either by increasing the number of individuals within clusters or by increasing the number of clusters. Increasing the number of clusters will usually enhance the generalisability of the study and will give greater flexibility at the time of analysis,¹⁵ but the relative cost of increasing the number of clusters in the study, rather than the number of individuals within clusters, will also be an important consideration.

(6) Consider the use of matching or stratification of clusters where appropriate

Stratification entails assigning clusters to strata classified according to cluster level prognostic factors. Equal numbers of clusters are then allocated to each intervention group from within each stratum. Some stratification or matching will often be necessary in area based or organisation based evaluations because simple randomisation will not usually give balanced intervention groups when a small number of clusters is randomised. However, stratification is useful only when the stratifying factor is fairly strongly related to the outcome.

The simplest form of stratified design is the matched pairs design, in which each stratum contains just two clusters. We advise caution in the use of the matched pairs design for two reasons. Firstly, the range of analytical methods appropriate for the matched design is more limited than for studies which use unrestricted allocation or stratified designs in which several clusters are randomised to each intervention group within strata.¹⁶ Secondly, when the number of clusters is less than about 20, a matched analysis may have less statistical power than an unmatched analysis.¹⁷ If matching is thought to be essential at the design stage, an unmatched cluster level analysis is worth considering.¹⁸ Stratified designs in which there are four or more clusters per stratum do not suffer from the limitations of the paired design.

(7) Consider different approaches to repeated assessments in prospective evaluations

Two basic sampling designs may be used for follow up: the cohort design, in which the same subjects from the study clusters are used at each measurement occasion, and the repeated cross sectional design, in which a fresh sample of subjects is drawn from the clusters at each measurement occasion.^{19,20} The cohort design is more appropriate when the focus of the study is on the effect of the programme at the level of the individual subject. The repeated cross sectional design, on the other hand, is more appropriate when the focus of interest is a cluster level index of health such as disease prevalence. The cohort design is potentially more powerful than the repeated cross sectional design because repeated observations on the same individuals tend to be correlated over time and may be used to reduce the variation of the estimated intervention effect. However, the repeated cross sectional design is more likely to give results that are representative of the clusters at the later measurement occasions, particularly for studies with long follow up.

(8) Allow for clustering at the time of analysis

Standard statistical methods are not appropriate for the analysis of individual level data from organisation based evaluations because they assume that the responses of different subjects are independent.² Standard methods may underestimate the standard error of the intervention effect, resulting in confidence intervals that are too narrow and P values that are too small.

Outcomes can be compared between intervention groups at the level of the cluster, applying standard statistical methods to the cluster means or proportions, or at the level of the individual, using formulas that have

been adjusted to allow for the similarity between individuals.²

Individual level analyses allow for the similarity between individuals within the same cluster, by incorporating the design effect into conventional standard error formulas that are used for hypothesis testing and estimating confidence intervals.^{2,21} For adjusted individual level analyses the intraclass correlation coefficient can be estimated from the study data in order to calculate the design effect. About 20-25 clusters are required to estimate the intraclass correlation coefficient with a reasonable level of precision and a cluster level analysis is to be preferred when there are fewer clusters than this.

(9) Allow for confounding at both individual and cluster levels

When confounding variables need to be controlled for at individual level or the cluster level, regression methods for clustered data should be used. The method of generalised estimating of equations treats the dependence between individual observations as a nuisance factor and provides estimates that are corrected for clustering. Random effects models (multilevel models) explicitly model the association between subjects in the same cluster. These methods may be used to estimate intervention effects, controlling for both individual level and cluster level characteristics.^{22,23} Regression methods for clustered data require a fairly large number of clusters but may be used with clusters that vary in size.

(10) Include estimates of intraclass correlation and components of variance in published reports

To aid the planning of future studies, researchers should publish estimates of the intraclass correlation for key outcomes of interest, for different types of subjects, and for different levels of geographical and organisational clustering.¹²⁻¹⁴

Recommendations

Investigators will need to consider the circumstances of their own evaluation and use discretion in applying these guidelines to specific circumstances. Conducting cluster based evaluations may present unusual difficulties. The issue of informed consent needs careful consideration.²⁴ Interventions and data management within clusters should be standardised, and the delivery of the intervention should usually be monitored through the collection of both qualitative and quantitative information, which may help to interpret the outcome of the study.

This article is adapted from *Health Services Research Methods: A Guide to Best Practice*, edited by Nick Black, John Brazier, Ray Fitzpatrick, and Barnaby Reeves, published by BMJ Books. We thank Kate Hann for commenting on the manuscript.

Funding: This work was supported by a contract from the NHS R&D Health Technology Assessment Programme. The views expressed do not necessarily reflect those of the NHS Executive.

Competing interests: None declared.

1 Murray DM. *Design and analysis of group randomised trials*. New York: Oxford University Press, 1998.

2 Donner A, Klar N. Cluster randomisation trials in epidemiology: theory and application. *J Stat Planning Inference* 1994;42:37-56.

- 3 Donner A, Brown KS, Brasher P. A methodological review of non-therapeutic intervention trials employing cluster randomisation, 1979-1989. *Int J Epidemiol* 1990;19:795-800.
- 4 McKinlay JB. More appropriate evaluation methods for community-level health interventions. *Evaluation Review* 1996;20:237-43.
- 5 Whiting-O'Keefe QE, Henke C, Simborg DW. Choosing the correct unit of analysis in medical care experiments. *Med Care* 1984;22:1101-14.
- 6 Grosskurth H, Mosha F, Todd J, Mwijarubi E, Klokke E, Seikoto K, et al. Improved treatment of sexually transmitted diseases on HIV infection in rural Tanzania: randomised controlled trial. *Lancet* 1995;346:530-6.
- 7 Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;312:1215-8.
- 8 Cook TD, Campbell DT. *Quasi-experimentation. Design and analysis issues for field settings*. Chicago: Rand McNally, 1979.
- 9 Donner A, Birkett N, Buck C. Randomisation by cluster. Sample size requirements and analysis. *Am J Epidemiol* 1981;114:906-14.
- 10 Donner A. Sample size requirements for cluster randomisation designs. *Stat Med* 1992;11:743-50.
- 11 Hsieh FY. Sample size formulae for intervention studies with the cluster as unit of randomisation. *Stat Med* 1988;7:1195-201.
- 12 Hannan PJ, Murray DM, Jacobs DR Jr, McGovern PG. Parameters to aid in the design and analysis of community trials: intraclass correlations from the Minnesota Heart Health Programme. *Epidemiology* 1994;5:88-95.
- 13 Ukoumunne OC, Gulliford MC, Chinn S, Sterne JAC, Burney PGJ. Methods for evaluating area-wide and organisation-based interventions in health and health care. *Health Technol Assess* 1999;3(5).
- 14 Gulliford MC, Ukoumunne OC, Chinn S. Components of variance and intraclass correlations for the design of community-based surveys and intervention studies: data from the health survey for England 1994. *Am J Epidemiol* 1999;149:876-83.
- 15 Thompson SG, Pyke SDM, Hardy RJ. The design and analysis of paired cluster randomised trials: an application of meta-analysis techniques. *Stat Med* 1997;16:2063-79.
- 16 Klar N, Donner A. The merits of matching: a cautionary tale. *Stat Med* 1997;16:1753-6.
- 17 Martin DC, Diehr P, Perrin EB, Koepsell TD. The effect of matching on the power of randomised community intervention studies. *Stat Med* 1993;12:123-31.
- 18 Diehr P, Martin DC, Koepsell T, Cheadle A. Breaking the matches in a paired t-test for community interventions when the number of pairs is small. *Stat Med* 1995;14:1491-504.
- 19 Feldman HA, McKinlay SM. Cohort versus cross-sectional design in large field trials: precision, sample size and a unifying model. *Stat Med* 1994;13:61-78.
- 20 Diehr P, Martin DC, Koepsell T, Cheadle A, Psaty BM, Wagner EH. Optimal survey design for community intervention evaluations: cohort or cross-sectional? *J Clin Epidemiol* 1995;48:1461-72.
- 21 Donner A, Klar N. Confidence interval construction for effect measures arising from cluster randomisation trials. *J Clin Epidemiol* 1993;46:123-31.
- 22 Rice N, Leyland A. Multi-level models application to health data. *J Health Serv Res Policy* 1996;1:154-64.
- 23 Duncan C, Jones K, Moon G. Context, composition and heterogeneity: using multi-level models in health research. *Soc Sci Med* 1998;46:97-117.
- 24 Edwards SJL, Lilford RJ, Braunholtz DA, Jackson JC, Hewison J, Thornton J. Ethics of randomised trials. In: Black N, Brazier J, Fitzpatrick R, Reeves B, eds. *Health services research methods. A guide to best practice*. London: BMJ Books, 1998:98-107.



Available from the BMJ bookshop (www.bmjbookshop.com)

Corrections and clarifications

Risk factors for human hantavirus infection: Franco-Belgian collaborative case-control study during 1995-6 epidemic

In this paper by N S Crowcroft and colleagues (26 June, p 1737-8) the names of two authors were transposed in the list of addresses. J-C Desencos is head of the infectious diseases unit at the Réseau National de Santé Publique, Saint-Maurice, France; and F Van Loock is an epidemiologist at the Scientific Institute of Public Health (Louis Pasteur) in Brussels, Belgium.

Annual general meeting of the BMA

In this letter by David Gullick (3 July, p 59) the second sentence of the first paragraph was misleading. It should have started: "It will be proposed that our 4000-odd overseas members (except those in the armed forces)..."

Obituaries

Dr Gordon Cunningham Taylor (19 June, p 1702) was incorrectly described as a lieutenant general in the Royal Army Medical Corps. He was a lieutenant colonel.

In the obituary of Dr Kevin Anthony Valiant (24 July, p 262), Dr Valiant's surname was incorrectly spelt.

takes an interdisciplinary approach to stemming the epidemic, building links with other international organisations, non-governmental organisations, the private sector, and the research community.

Communicating reality and vision

Much remains to be done to raise awareness and concern about cancer in the developing world. The yawning gap between poor and rich countries persists, and cheap effective technologies such as hepatitis B vaccine are not applied. There is a pressing need to deal pragmatically with today's problems by setting realistic priorities. Yet health professionals also have a responsibility to expand what is feasible. As Article 27 of the Universal Declaration of Human Rights states, "Everyone has the right . . . to share in scientific advancement and its benefits." This vision can be communicated through persuasive and practical arguments for placing cancer in developing countries squarely in context—and firmly on the agenda.

I thank N Muñoz and R Sankaranarayan for discussions and advice, and P Kleihues, DM Parkin, K Sikora, A Narinesingh, E

LeGresley, and Y Daikh for their comments. Special thanks go to KJ Hughes.

Competing interests: None declared.

- 1 UNAIDS/WHO. *Epidemiological fact sheet on HIV/AIDS and sexually transmitted diseases*. Geneva: UNAIDS/WHO, 1998.
- 2 World Health Organisation Programme on Cancer Control. *Cancer strategies for the new millennium*. Proceedings of a conference held at the Royal College of Physicians, London, 1998. (www.who-pcc.iarc.fr/Publications/Publications.html; accessed 2 March 1999.)
- 3 Murray CJL, Lopez AD, eds. *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*. Cambridge, MA: Harvard University Press, 1996.
- 4 Pisani P, Parkin DM, Muñoz N, Ferlay J. Cancer and infection: estimates of the attributable fraction in 1995. *Cancer Epidemiol Biomarkers Prev* 1997;6:387-400.
- 5 Parkin DM, Pisani P, Ferlay J. Estimates of the worldwide incidence of 25 major cancers in 1990. *Int J Cancer* 1999;80:827-41.
- 6 Pisani P, Parkin DM, Bray F, Ferlay J. Estimates of the worldwide mortality from 25 major cancers in 1990. *Int J Cancer* 1999 (in press).
- 7 Sankaranarayanan R, Black RJ, Parkin MD, eds. *Cancer survival in developing countries*. Lyons: International Agency for Research on Cancer, 1998.
- 8 Walboomers JMM, Jacobs MV, Manos MM, Bosch FX, Kummer JA, Shah KV, et al. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J Pathol* 1999 (in press).
- 9 WHO Expanded Programme on Immunisation. *Hepatitis B vaccine—making global progress*. Geneva: WHO, 1997.

(Accepted 9 April 1999)

Methods in health service research

An introduction to bayesian methods in health technology assessment

David J Spiegelhalter, Jonathan P Myles, David R Jones, Keith R Abrams

This is the third of four articles

MRC Biostatistics Unit, Institute of Public Health, Cambridge CB2 2SR

David J Spiegelhalter, senior statistician

Jonathan P Myles, research assistant

Department of Epidemiology and Public Health, University of Leicester, Leicester LE1 6TP

Keith R Abrams, senior lecturer in medical statistics

David R Jones, professor of medical statistics

Correspondence to: Dr Spiegelhalter david.spiegelhalter@

mrc-bsu.cam.ac.uk

Series editor: Nick Black

BMJ 1999;319:508-12

Bayes's theorem arose from a posthumous publication in 1763 by Thomas Bayes, a non-conformist minister from Tunbridge Wells. Although it gives a simple and uncontroversial result in probability theory, specific uses of the theorem have been the subject of considerable controversy for more than two centuries. In recent years a more balanced and pragmatic perspective has emerged, and in this paper we review current thinking on the value of the Bayesian approach to health technology assessment.

A concise definition of bayesian methods in health technology assessment has not been established, but we suggest the following: the explicit quantitative use of external evidence in the design, monitoring, analysis, interpretation, and reporting of a health technology assessment. This approach acknowledges that judgments about the benefits of a new technology will rarely be based solely on the results of a single study but should synthesise evidence from multiple sources—for example, pilot studies, trials of similar interventions, and even subjective judgments about the generalisability of the study's results.

A bayesian perspective leads to an approach to clinical trials that is claimed to be more flexible and ethical than traditional methods,¹ and to elegant ways of handling multiple substudies—for example, when simultaneously estimating the effects of a treatment on many subgroups.² Proponents have also argued that a bayesian approach allows conclusions to be provided in a form that is most suitable for decisions specific to patients and decisions affecting public policy.³

Summary points

Bayesian methods interpret data from a study in the light of external evidence and judgment, and the form in which conclusions are drawn contributes naturally to decision making

Prior plausibility of hypotheses is taken into account, just as when interpreting the results of a diagnostic test

Scepticism about large treatment effects can be formally expressed and used in cautious interpretation of results that seem "too good to be true"

Multiple subanalyses can be brought together by formally expressing a belief that their conclusions should be broadly similar

Use of bayesian methods in health technology assessment should be pursued cautiously; guidelines, software, and critically evaluated case studies are needed

Many questions remain: notably, to what extent the scientific community or regulatory authorities will allow the explicit consideration of evidence that is not totally derived from observed data. In this article we

outline the available literature, discuss the main techniques that are being suggested, and provide some recommendations for future work.

Nature of the evidence

A “bayesian” approach can be applied to many scientific issues, and a search for this term in the Institute for Scientific Information’s database yielded nearly 4000 papers over the period 1990-8. About 200 of these were relevant to health technology assessment. Using these as a source for forward and backward searches, and searching other databases (Embase and Medline) and sources, we identified about 300 papers, including about 30 reports of studies taking a fully bayesian perspective. A considerable further number of studies have taken a so called “empirical Bayes” approach, which uses elements of bayesian modelling without giving a bayesian interpretation to the conclusions; these are further mentioned below.

The published studies are dispersed throughout the literature and, apart from one recent collection of papers,⁴ the only textbook which might be considered to be on bayesian methods in health technology assessment focuses on the confidence profile approach.⁵ Published studies are mainly demonstrations of the approach rather than complete assessments, and though many articles advocate bayesian methods, practical take-up seems low.

Findings

Philosophy of the bayesian approach

Bayes’s theorem is a formula that shows how existing beliefs, formally expressed as probability distributions, are modified by new information. Diagnostic testing is a familiar situation to which the theorem can be applied; a doctor’s prior belief about whether a patient has a particular disease (based on knowledge of the prevalence of the disease in the community and the patient’s symptoms) will be modified by the result of the test.⁶

The unknown piece of information may, however, be a somewhat more intangible quantity than an individual’s true diagnosis—for example, the average survival benefit of drug A over drug B in a particular group of patients. Such quantities are not directly observable in any reasonably sized experiment and are considered to be unknown variables. Just as the full evaluation of a diagnostic test requires the prevalence of the disease to be specified, a bayesian analyst is prepared to make the bold step of specifying a probability distribution expressing the relative plausibility for this unknown quantity, before taking into account any evidence from a study. This “prior” distribution can then be combined with evidence from the study to form a “posterior” (formally proportional to the product of the prior and the likelihood function). The box shows an example.

The posterior distribution provides probabilities of events of clinical interest and so one could say, for example, that under specified assumptions “the chance is 15% that drug A improves average survival by at least three months over drug B.” This type of statement is impossible to make within the traditional statistical framework, in which the interpretation of P values and confidence intervals depends on rather convoluted

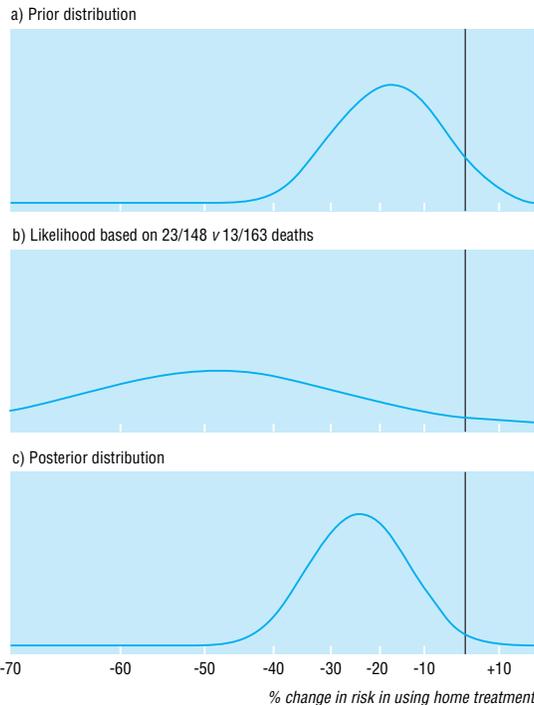


Fig 1 Prior (a), likelihood (b), and posterior (c) distributions arising from reanalysis by Pocock and Spiegelhalter⁷ of the GREAT trial of home thrombolysis.⁸ The prior distribution represents a summary of evidence external to the trial, the likelihood expresses evidence from the trial itself, and the posterior distribution pools these two sources by multiplying the two curves together

statements concerning the long run properties of statistical procedures under null hypotheses.

The table briefly summarises some major distinctions between the bayesian and the traditional

Bayes’s theorem after a randomised trial

Pocock and Spiegelhalter⁷ discuss a small trial of early thrombolytic treatment in preventing deaths from myocardial infarction, which had reported a remarkable 49% reduction in mortality.⁸ On the basis of both published and unpublished large trials, they argued that if treatment were provided two hours earlier “a 15-20% reduction in mortality is highly plausible, while the extremes of no benefit and a 40% reduction are both unlikely.” This opinion could be represented as a prior distribution as shown in figure 1(a), which expresses the relative plausibility arising from this external evidence.

Figure 1(b) shows the “likelihood” for the true risk reduction arising from the trial itself, which is simply proportional to the chance of observing the data (23/148 deaths in controls v 13/163 deaths with active treatment) for each hypothesised risk reduction. Bayes’s theorem states that the two sources of evidence can be combined by multiplying the prior and likelihood curves together and then making the total area under the resulting curve be equal to 1—this is the “posterior” distribution and is shown in figure 1(c). The evidence in the likelihood has been pulled back towards the prior opinion, thus formally representing the suspicion that the trial results were “too good to be true.”

The resulting distribution provides an easily interpretable summary of the total evidence, and posterior probabilities for hypotheses of interest can then be read from the graph. For example, the most likely benefit is a reduction in risk of around 24% (half that observed in the trial), the posterior probability that the risk is reduced by at least 50% is only 5%, and a 95% confidence interval is from 43% to 0% risk reduction. Subsequent experience has reinforced the conclusion of this analysis that it is very unlikely that home thrombolysis reduces mortality by 50%.

Brief comparison of bayesian and frequentist methods in randomised trials

Issue	Frequentist methods	Bayesian methods
Prior information other than that in the study being analysed	Informally used in design	Used formally by specifying a prior probability distribution
Interpretation of the parameter of interest	A fixed state of nature	An unknown quantity which can have a probability distribution
Basic question	"How likely is the data, given a particular value of the parameter?"	"How likely is a particular value of the parameter given the data?"
Presentation of results	Likelihood functions, P values, confidence intervals	Plots of posterior distributions of the parameter, calculation of specific posterior probabilities of interest, and use of the posterior distribution in formal decision analysis
Interim analyses	P values and estimates adjusted for the number of analyses	Inference not affected by the number or timing of interim analyses
Interim predictions	Conditional power analyses	Predictive probability of getting a firm conclusion
Dealing with subsets in trials	Adjusted P values (for example, Bonferroni)	Subset effects shrunk towards zero by a "sceptical" prior

approach. The latter is sometimes termed "frequentist" as it is based on the long run frequency properties of statistical procedures. There are many papers summarising the bayesian philosophy and its application to randomised trials: Cornfield's is a notable early example,⁹ and other authors have argued for the flexibility, coherence, and intuitiveness of the approach.^{1-3 10} Several authors have highlighted how the bayesian approach leads naturally into a formal decision theoretical approach to randomised trials.¹¹

Quantifying prior beliefs

The bayesian approach is most controversial when there is no hard evidence for the prior distribution and we have to rely on subjective judgment. This considerably broadens the area of potential application, although the reasonableness of the judgments will need to be justified. The traditional terms prior and posterior may also be misleading, giving the impression that the prior has to be fixed before the evidence is examined. It is more helpful to think of the prior as summarising all external evidence about the quantity of interest—for example, other published studies—which might arise during or after the study that is being considered.

One source of a prior distribution is the pooled subjective opinion of informed experts, which can be elicited interactively by using computer programs¹² or questionnaire methods.¹³ Such opinions should rely on extensive experience: for example, Peto and Baigent state that "it is generally unrealistic to hope for large treatment effects" but that "it might be reasonable to hope that a new treatment for acute stroke or acute myocardial infarction could reduce recurrent stroke or death in hospital from 10% to 9% or 8% ... but not to hope that it could halve in-hospital mortality."¹⁴ This closely mimics the prior opinion used in the box above to illustrate how extreme results based on small studies should not be taken at face value. Another source of prior opinions is, of course, meta-analyses of previous similar studies.

One important use of a prior distribution is in planning the sample size of a randomised trial. Instead of using a single (possibly optimistic) alternative hypothesis as the basis for the power calculation, the prior distribution can be used to produce an "expected power," taking into account reasonable uncertainty about the true treatment effect.¹³

There has been an increasing move towards "off the shelf" priors—for example, those intended to represent the opinions of an archetypal "sceptic" and those of an "enthusiast"¹⁵; these can be used to represent extreme opinions in sensitivity analyses and

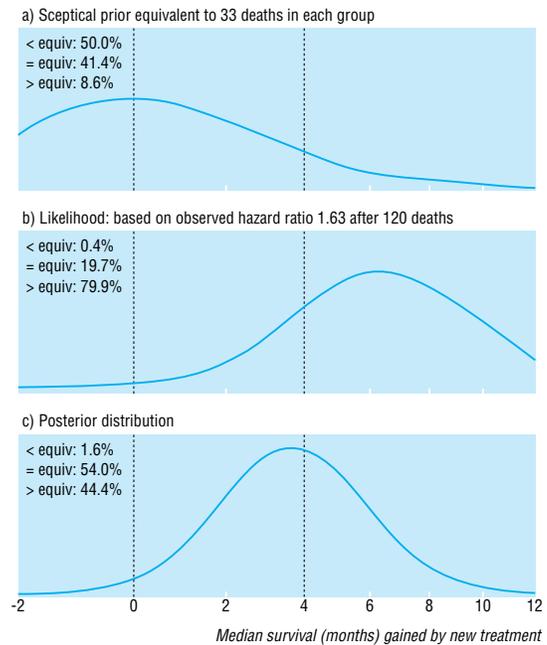


Fig 2 Prior, likelihood, and posterior distributions arising from Cancer and Leukaemia Group B trial of standard radiotherapy versus additional chemotherapy in advanced lung cancer.¹⁵ Dashed lines give boundaries of range of clinical equivalence, taken to be 0 and 4 months median improvement in survival. Numbers by each graph show probabilities of lying below, within, and above the range of equivalence

in sequential monitoring of trials (see below). One published example concerns the use of sceptical priors in determining whether there is sufficient evidence for a treatment to be generally recommended (box).

Applications in monitoring randomised trials

In the traditional frequentist approach, randomised trials are designed to have a fixed chance (usually 5%) of incorrectly rejecting the null hypothesis, and various techniques have been developed for adjusting the apparent significance level of a result to allow for the fact that the data have been analysed more than once. The bayesian approach sees no need for this and instead monitors the trial on the basis of the current posterior distribution, providing an updated summary of the evidence about the treatment effect at the time of any analysis. Several monitoring schemes have been suggested, some of which are based on decision theory.¹¹ The most frequently illustrated technique is simply based on the "tail" areas of the posterior distribution—

for example, stop the trial if the chance that the treatment is more effective than control is greater than 99%.¹⁷ If desired, the probability of the treatment effect being greater than some clinically important difference may be used, or, in the case of equivalence studies, that the treatment difference is less than, say, 10%.

A sceptical prior may be thought of as a handicap that the trial data must overcome in order to provide convincing evidence of benefit. In the light of early positive results, the approach shows a degree of conservatism which can be remarkably similar to that of frequentist stopping rules.¹⁸ The use of sceptical priors has been described in a tutorial and in meta-analyses,^{19, 20} and a senior statistician with the US Food and Drug Administration has said that he “would like to see [sceptical priors] applied in more routine fashion to provide insight into our decision making.”²¹

The table also considers predictions made at an interim stage in a randomised trial. Whereas the frequentist conditional power calculations are based on a hypothesised value of the true treatment effect, a bayesian approach can answer a crucial question: if we continue the study, what is the chance we will get a significant result?

Multiplicity—estimating the prior

We often wish simultaneously to carry out a set of related analyses—for example, meta-analysis of individual trial results—allowing for between centre variability in the analysis of a multicentre trial or analysing subsets of cases in a single trial. We call these subanalyses. The traditional frequentist approach tries to maintain a constant probability of wrongly rejecting the null hypothesis (type I error) by some adjustment—for example, a Bonferroni method for multiple comparisons.

The bayesian approach integrates subanalyses by assuming that the unknown quantities (for example, the treatment effects specific to subsets) have a common prior distribution, with the important difference that this prior distribution has unknown parameters that need to be estimated. Such models are known as hierarchical and can, in theory, have any number of levels, although three is generally enough. Non-bayesian versions (multi-level, random effects and random coefficient models) use either likelihood or “empirical Bayes” approaches to estimate the model parameters.

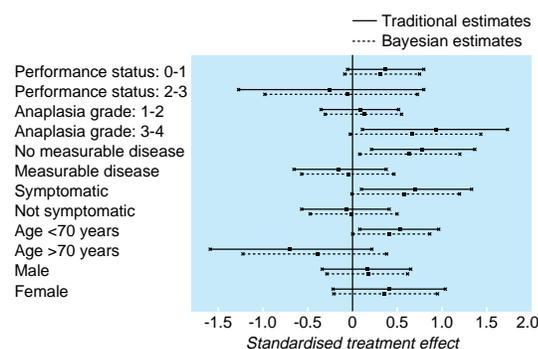


Fig 3 Traditional and bayesian estimates of standardised treatment effects in a randomised trial of treatments for cancer. The bayesian estimates are pulled towards the overall treatment effect by a degree determined by the empirical heterogeneity of the subset results

Is a confirmatory trial necessary?

Parmar et al illustrate the use of a sceptical prior distribution in deciding whether or not to perform a confirmatory randomised trial.¹⁶ They discuss a Cancer and Leukaemia Group B trial of radiotherapy and chemotherapy versus standard radiotherapy in patients with locally advanced stage III non-small cell lung cancer. This trial showed an adjusted median improvement in survival of 6.3 months (95% confidence interval 1.4 to 13.3 months) in favour of the new treatment, which has a two sided P value of 0.008. They give two reasons why this might not lead to an immediate recommendation for radiotherapy and chemotherapy as standard treatment. Firstly, the toxicity of chemotherapy might mean a minimum worthwhile improvement is demanded; the authors suggest a figure of around four months. Secondly, a natural scepticism exists about new cancer treatments, derived from long experience of failed innovations.

These two aspects can be formalised within the bayesian framework. Firstly, one can report the probability that the new treatment not only provides a positive improvement but that this exceeds a minimum clinically worthwhile improvement. Secondly, scepticism is expressed by a prior distribution that is centred on zero improvement and shows a 5% chance that the true improvement is greater than the alternative hypothesis in this study—namely, that the true improvement is five months.

Figure 2 shows this sceptical prior distribution, which is equivalent evidence to that of an “imaginary” trial in which 33 patients taking each treatment died. The dashed vertical lines indicate the null hypothesis of no improvement and the minimum clinically worthwhile improvement of four months. Between these lie what can be termed the range of equivalence, and the figure shows that the sceptical prior expresses a probability of 41% that the true benefit lies in the range of equivalence and only 9% that the new treatment is clinically superior.

The likelihood function shows the inferences to be made from the data alone, assuming a “uniform” prior on the range of possible improvements; Parmar et al call this an enthusiastic prior. The probability that the new treatment is actually inferior is 0.4% (equivalent to the one sided P value of $0.008 \div 2$). The probability of clinical superiority is 80%, which might be considered sufficient to change treatment policy.

The posterior distribution shows the impact of the sceptical prior, in that the chance of clinical superiority is reduced to 44%, hardly sufficient to change practice. In fact, Parmar et al report that the National Cancer Institute intergroup trial investigators were unconvinced by the Cancer and Leukaemia Group B trial due to their previous negative experience, and so carried out a further study. They found a significant median improvement, but of only 2.4 months, suggesting that the sceptical approach might have given a more reasonable estimate.

By assuming a common prior distribution for each subanalysis we are expressing scepticism about large differences in their outcomes, although the precise degree of similarity is generally considered unknown and estimated from the data—for example, by measuring the between trial variability in a meta-analysis. Full bayesian and empirical Bayes approaches can lead to similar conservatism (box).²²

Non-randomised studies and synthesis of evidence

Most authors have concentrated on the application of bayesian methods when designing randomised trials or pooling results from published trials, but a small number of papers have considered applying these methods to data collected from non-randomised studies. For example, in a paper analysing data from two case-control studies (one being very small) and a cohort study, the authors show the results of using different sources of information for the prior and likelihood.²⁴ Other authors have discussed the integration of evidence from several types of non-randomised studies²⁵ and the integration of findings from both randomised and non-randomised studies within a bayesian framework.²⁶

Bayes's theorem for subset analysis

Dixon and Simon describe a bayesian approach to dealing with subset analysis in a randomised trial in advanced colorectal cancer.²⁵ The solid horizontal lines in figure 3 show the standardised treatment effects within a range of subgroups, using traditional methods for estimating treatment by subgroup interactions. Four of the 12 intervals exclude zero; because multiple hypotheses are being tested, however, an adjustment technique such as Bonferroni might be used to decrease the apparent statistical significance of these findings.

The bayesian approach is to assume that deviations from the overall treatment effect that are specific to subgroups have a prior distribution centred at zero but with an unknown variability; this variability is then given its own prior distribution. Since the degree of scepticism is governed by the variance of the prior distribution, the observed heterogeneity of treatment effects between subgroups will influence the degree of scepticism being imposed.

The resulting bayesian estimates are shown as dashed lines in figure 3. They tend to be pulled towards each other, owing to the prior scepticism about substantial interaction effects between subgroups and treatments. Only one 95% confidence interval now excludes zero, that for the subgroup with no measurable metastatic disease. Dixon and Simon mention that this was the conclusion of the original trial; the bayesian analysis has the advantage of not relying on somewhat arbitrary adjustment techniques as it can be generalised to any number of subsets, and it provides a unified means of both providing estimates and tests of hypotheses.

Decision making

Another important feature of a bayesian approach is the way in which the resulting posterior probability distribution can be combined with quantitative measures of utility as part of a formal decision analysis. As with the elicitation of beliefs regarding probabilities, the elicitation and quantification of utilities is challenging, and this is one of the least developed areas of bayesian analysis. Such formal uses of decision theory have been applied in health technology assessments in various settings, including the development of clinical recommendations for prevention of stroke,²⁷ monitoring and analysis in randomised trials,¹¹ and assessment of environmental contamination on public health.²⁸

Recommendations

Bayesian analysis is widely used in a variety of non-medical fields, including engineering, image processing, expert systems, decision analysis, gene sequencing, financial predictions, and neural networks, and increasingly in complex epidemiological models. Health technology assessment has been slow to adopt bayesian methods; this could be due to a reluctance to use prior opinions, unfamiliarity, mathematical complexity, lack of software, or conservatism of the health care establishment and, in particular, the regulatory authorities.

There are strong philosophical reasons for using a bayesian approach, but the current literature emphasises the practical advantages in handling complex interrelated problems and in making explicit and accountable what is usually implicit and hidden, thereby clarifying discussions and disagreements. Perhaps the most persuasive reason is that the analysis tells us what we want to know: how should this piece of evidence change what we currently believe?

The perceived problems with the bayesian approach largely concern the source of the prior and the interpretations of the conclusions. There are also practical difficulties in implementation and software. Current international guidelines for statistical sub-

missions to drug regulatory authorities state that "the use of bayesian and other approaches may be considered when the reasons for their use are clear and when the resulting conclusions are sufficiently robust,"²⁹ and it seems sensible that experience should be gained in the use of bayesian approaches in health technology assessment in parallel with traditional approaches, with careful consideration of the sensitivity of results to prior distributions.

For future practical and methodological developments, we recommend:

- An extended set of case studies showing practical aspects of the bayesian approach, in particular for prediction and handling multiple substudies, in which mathematical details are minimised;
- The development of standards for the performance and reporting of bayesian analyses;
- The development and dissemination of software for bayesian analysis, preferably as part of existing programs.

This article is adapted from *Health Services Research Methods: A Guide to Best Practice*, edited by Nick Black, John Brazier, Ray Fitzpatrick, and Barnaby Reeves, published by BMJ Books.

Competing interests: None declared.

- 1 Kadane JB. Prime time for Bayes. *Contr Clin Trials* 1995;16:313-8.
- 2 Breslow N. Biostatistics and Bayes. *Stat Sci* 1990;5:269-84.
- 3 Lilford RJ, Braunholtz D. The statistical basis of public policy—a paradigm shift is overdue. *BMJ* 1996;313:603-7.
- 4 Berry DA, Stangl DK. *Bayesian biostatistics*. New York: Dekker, 1996.
- 5 Eddy DM, Hasselblad V, Shachter R. *Meta-analysis by the confidence profile method: the statistical synthesis of evidence*. San Diego, CA: Academic, 1992.
- 6 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine*, 2nd ed. Boston: Little Brown, 1991.
- 7 Pocock S, Spiegelhalter DJ. Domiciliary thrombolysis by general practitioners. *BMJ* 1992;305:1015.
- 8 GREAT Group. Feasibility, safety and efficacy of domiciliary thrombolysis by general practitioners: Grampian region early anistreplase trial. *BMJ* 1992;305:548-53.
- 9 Cornfield J. Recent methodological contributions to clinical trials. *Am J Epidemiol* 1976;104:408-21.
- 10 Lewis RJ, Wears RL. An introduction to the Bayesian analysis of clinical trials. *Ann Emergency Med* 1993;22:1328-36.
- 11 Berry DA, Wolff MC, Sack D. Decision-making during a phase-III randomised controlled trial. *Contr Clin Trials* 1994;15:360-78.
- 12 Chaloner K, Church T, Louis TA, Matts JP. Graphical elicitation of a prior distribution for a clinical-trial. *Statistician* 1993;42:341-53.
- 13 Parmar MKB, Spiegelhalter DJ, Freedman LS. The chart trials: bayesian design and monitoring in practice. *Stat Med* 1994;13:1297-312.
- 14 Peto R, Baigent C. Trials: the next 50 years. *BMJ* 1998;317:1170-1.
- 15 Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomised trials. *J Roy Stat Soc Series A* 1994;157:357-87.
- 16 Parmar MKB, Ungerleider RS, Simon R. Assessing whether to perform a confirmatory randomised clinical trial. *J Natl Cancer Inst* 1996;88:1645-51.
- 17 Freedman LS, Spiegelhalter DJ, Parmar MKB. The what, why and how of bayesian clinical trials monitoring. *Stat Med* 1994;13:1371-83.
- 18 Grossman J, Parmar MKB, Spiegelhalter DJ, Freedman LS. A unified method for monitoring and analysing controlled trials. *Stat Med* 1994;13:1815-26.
- 19 Fayers PM, Ashby D, Parmar MKB. Bayesian data monitoring in clinical trials. *Stat Med* 1997;16:1413-30.
- 20 Dersimonian R. Meta-analysis in the design and monitoring of clinical trials. *Stat Med* 1996;15:1237-48.
- 21 O'Neill R. Early stopping rules workshop: conclusions. *Stat Med* 1994;13:1493-9.
- 22 Louis TA. Using empirical Bayes methods in biopharmaceutical research. *Stat Med* 1991;10:811-29.
- 23 Dixon DO, Simon R. Bayesian subset analysis in a colorectal cancer clinical trial. *Stat Med* 1992;11:13-22.
- 24 Ashby D, Hutton J, McGee M. Simple bayesian analyses for case-control studies in cancer epidemiology. *Statistician* 1993;42:385-97.
- 25 Eddy DM, Hasselblad V, Shachter R. An introduction to a bayesian method for meta-analysis—the confidence profile method. *Medical Decision Making* 1990;10:15-23.
- 26 Abrams KR, Jones DR. Meta-analysis and the synthesis of evidence. *IMA J Math Med Biol* 1995;12:297-313.
- 27 Parmigiani G, Ancukiewicz M, Matchar D. Decision models in clinical recommendations development: the stroke prevention policy model. In: Berry A, Stangl DK, eds. *Bayesian biostatistics*. New York: Dekker, 1996:207-33.
- 28 Wolfson LJ, Kadane JD, Small MJ. Expected utility as a policy-making tool: an environmental health example. In: Berry A, Stangl DK, eds. *Bayesian biostatistics*. New York: Dekker, 1996:261-77.
- 29 International Conference on Harmonisation. Statistical principles for clinical trials, 1998. www.ich.org/pdf/pma/e9.pdf; accessed May 1999.

- 24 Hansson L, Lindholm LH, Niskanen L, Lanke J, Hedner T, Niklason A, et al. Effect of angiotensin-converting enzyme inhibition compared with conventional therapy on cardiovascular morbidity and mortality in hypertension: the captopril prevention project (CAPPP). *Lancet* 1999; 353:611-5.
- 25 SHEP Cooperative Research Group. Prevention of stroke by antihypertensive drug treatment in older persons with isolated systolic hypertension. Final results of the systolic hypertension in the elderly program (SHEP). *JAMA* 1991;265:3255-64.
- 26 Thijs L, Fagard R, Lijnen P, Staessen J, Van Hoof R, Amery A. A meta-analysis of outcome trials in elderly hypertensives. *J Hypertens* 1992;10:1103-9.
- 27 Messerli FH, Grossman E, Goldbourt U. Are β -blockers efficacious as first-line therapy for hypertension in the elderly? A systematic review. *JAMA* 1998;279:1903-7.
- 28 Medical Research Council's General Practice Research Framework. Thrombosis prevention trial: randomised trial of low-intensity oral anticoagulation with warfarin and low-dose aspirin in the primary prevention of ischaemic heart disease in men at increased risk. *Lancet* 1998;351:233-41.
- 29 Downs JR, Clearfield M, Weis S, Whitney E, Shapiro DR, Beere PA, et al. Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels. Results of AFCAPS/TextCAPS. *JAMA* 1998;279:1615-22.
(Accepted 11 August 1999)

Appendix

Material for patients

- Patient information booklet: "Understanding High Blood Pressure"
- Fact sheets:
 - Selfhelp measures
 - Antihypertensive drugs

Blood pressure measurement
Reducing dietary salt
Blood pressure and kidney disease

- Diet sheet: "Healthy Eating"

Available from the British Hypertension Society Information Service, Blood Pressure Unit, St George's Hospital Medical School, Cranmer Terrace, London SW17 0RE (tel: 0181 725 3412; fax: 0181 725 2959; www.bhsinfo.hyp.ac.uk (for information service); website: www.bhs.hyp.ac.uk)

Material for doctors

- *Blood Pressure Measurement—Recommendations of the British Hypertension Society*. 3rd edition, 1997. (Edited by E O'Brien et al; price £4.95.)
- BHS/BMJ. *Recommendations for Blood Pressure Measurement*. CD Rom, price £58.75.
Available from BMJ Publications or the BMJ Bookshop, BMA House, London WC1H 9JR (tel: 0171 383 6244; fax: 0171 383 6455; orders@bmjbookshop.com).
- The Joint British Societies' Cardiac Risk Assessor computer program and copies of the Joint British Societies coronary heart disease risk assessment chart can be downloaded from the British Hypertension Society website (www.bhs.hyp.ac.uk).

Methods in health service research

Handling uncertainty in economic evaluations of healthcare interventions

Andrew H Briggs, Alastair M Gray

The constant introduction of new health technologies, coupled with limited healthcare resources, has engendered a growing interest in economic evaluation as a way of guiding decision makers towards interventions that are likely to offer maximum health gain. In particular, cost effectiveness analyses—which compare interventions in terms of the extra or incremental cost per unit of health outcome obtained—have become increasingly familiar in many medical and health service journals.

Considerable uncertainty exists in regard to valid economic evaluations. Firstly, several aspects of the underlying methodological framework are still being debated among health economists. Secondly, there is often considerable uncertainty surrounding the data, the assumptions that may have been used, and how to handle and express this uncertainty. In the absence of data at the patient level sensitivity analysis is commonly used; however, a number of alternative methods of sensitivity analysis exist, with different implications for the interval estimates generated (see box). Finally, there is a substantial amount of subjectivity in presenting and interpreting the results of economic evaluations.

The aim of this paper is to give an overview of the handling of uncertainty in economic evaluations of healthcare interventions.³ It examines how analysts have handled uncertainty in economic evalua-

Summary points

Economic evaluations are beset by uncertainty concerning methodology and data

A review of 492 articles published up to December 1996 found that a fifth did not attempt any analysis to examine uncertainty

Only 5% of these studies reported some measure of cost variance

Closer adherence to published guidelines would greatly improve the current position

Use of a methodological reference case will improve comparability

This is the last of four articles

Health Economics Research Centre, Institute of Health Sciences, University of Oxford, Oxford OX3 7LF

Andrew H Briggs
training fellow

Alastair M Gray
reader

Correspondence to:
A H Briggs
andrewbriggs@his.ox.ac.uk
Series editor:
Nick Black

BMJ 1999;319:635-8

tion, assembled data on the distribution and variance of healthcare costs, and proposed guidelines to improve current practice. It is intended as a contribution towards the development of agreed guidelines for analysts, reviewers, editors, and decision makers.⁴⁻⁷

Box 1: Sensitivity analysis

Sensitivity analysis involves systematically examining the influence of uncertainties in the variables and assumptions employed in an evaluation on the estimated results. It encompasses at least three alternative approaches.¹

- *One way sensitivity analysis* systematically examines the impact of each variable in the study by varying it across a plausible range of values while holding all other variables in the analysis constant at their “best estimate” or baseline value.
- *Extreme scenario analysis* involves setting each variable to simultaneously take the most optimistic (pessimistic) value from the point of view of the intervention under evaluation in order to generate a best (worst) case scenario.

Of course, in real life the components of an evaluation do not vary in isolation nor are they perfectly correlated, hence it is likely that one way sensitivity analysis will underestimate, and extreme scenario analysis overestimate, the uncertainty associated with the results of economic evaluation.

- *Probabilistic sensitivity analysis*, which is based on a large number of Monte Carlo simulations, examines the effect on the results of an evaluation when the underlying variables are allowed to vary simultaneously across a plausible range according to predefined distributions. These probabilistic analyses are likely to produce results that lie between the ranges implied by one way sensitivity analysis and extreme scenario analysis, and therefore may produce a more realistic estimate of uncertainty.²

Nature of the evidence

A structured review examined the methods used to handle uncertainty in the empirical literature, and this was supplemented by a review of methodological articles on the specific topic of confidence interval estimation for cost effectiveness ratios. The first step in the empirical review was a search of the literature to identify published economic evaluations that reported results in terms of cost per life year or cost per quality adjusted life year (QALY). This form of study was chosen as the results of these studies are commonly considered to be sufficiently comparable to be grouped together and reported in cost effectiveness league tables.

Searches were conducted for all such studies published up to the end of 1996 using Medline, CINAHL, Econlit, Embase, the Social Science Citation Index, and the economic evaluation databases of the Centre for Reviews and Dissemination at York University and the Office of Health Economics and International Federation of Pharmaceutical Manufacturers' Association. Articles identified as meeting the search criteria were reviewed by using a form designed to collect summary information on each study, including the disease area, type of intervention, nature of the data, nature of the results, study design, and the methods used to handle uncertainty. This information was entered as keywords into a database to allow interrogation and cross referencing of the database by category.

This overall dataset was then used to focus on two specific areas of interest, using subsets of articles to perform more detailed reviews. Firstly, all British studies were identified and reviewed in detail, and

information on the baseline results, the methods underlying those results, the range of results representing uncertainty, and the number of previously published results quoted for purposes of comparison were entered on to a relational database. By matching results by the methods used in a retrospective application of a methodological “reference case” (box),⁵ a subset of results with improved comparability was identified, and a rank ordering of these results was then attempted. Where a range of values accompanied the baseline results, the implications of this uncertainty for the rank ordering was also examined.

Secondly, all studies that reported cost data at the patient level were identified and reviewed in detail with respect to how they had reported the distribution and variance of healthcare costs. Thirdly, and in parallel with the structured review, five datasets of patient level cost data were obtained and examined to show how the healthcare costs in those data were distributed and to elucidate issues surrounding the analysis and presentation of differences in healthcare cost.

Economic analyses are not simply concerned with costs, but also with effects, with the incremental cost effectiveness ratio being the outcome of interest in most economic evaluations. Unfortunately, ratio statistics pose particular problems for standard statistical methods. The review examines a number of proposed methods that have appeared in the recent literature for estimating confidence limits for cost effectiveness ratios (when patient level data are available).

Findings**Trends in economic evaluations**

A total of 492 articles published up to December 1996 were found to match the search criteria and were fully reviewed. The review found an exponential rate of increase in published economic evaluations over time and an increasing proportion reporting cost per QALY results. Analysis of the articles in terms of the method used by analysts to handle uncertainty shows that the vast majority of studies (just over 70%) used one way sensitivity analysis methods to quantify uncertainty (see box 1). Of some concern is that almost 20% of studies did not attempt any analysis to examine uncertainty, although there is weak evidence to show that this situation has improved over time.

The “reference case”

The Panel on Cost-Effectiveness in Health and Medicine, an expert committee convened by the US Public Health Service in 1993, proposed that all published cost effectiveness studies contain at least one set of results based on a standardised set of methods and conventions—a reference case analysis—which would aid comparability between studies. The features of this reference case were set out in detail in the panel's report.⁵

The current review used this concept retrospectively, selecting for comparison a subset of results which conformed to the following conditions:

- An incremental analysis was undertaken;
- A health service perspective was employed; and
- Both costs and health outcomes were discounted at the UK Treasury approved rate of 6% per annum.

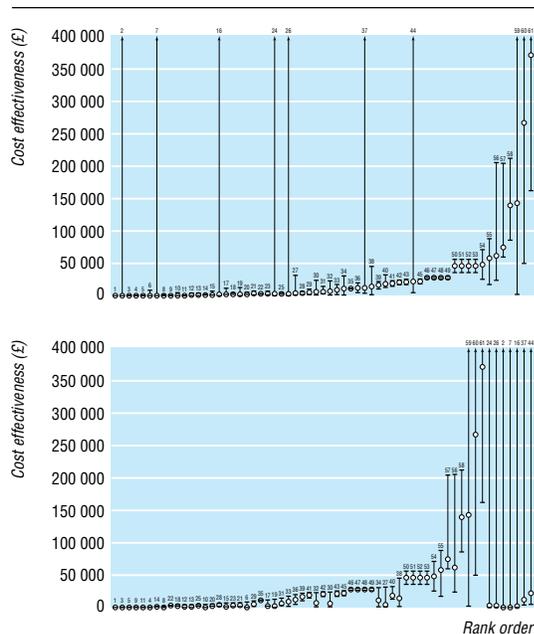


Fig 1 Alternative rank orderings of 61 British cost effectiveness results by baseline value (above) and highest sensitivity analysis value (below)

Handling of uncertainty

Of the 492 studies, 60 reported results for the United Kingdom. From these, 548 baseline results were extracted for different subgroups. The importance of separate baselines for different subgroups of patients is shown in the results of an evaluation of an implantable cardioverter defibrillator where the average cost per life year saved across the whole patient group—£57 000—masks important differences between patients with different clinical characteristics.⁸ For patients with a low ejection fraction and inducible arrhythmia that is not controlled by drugs, the cost effectiveness of the device is £22 000 per year of life saved. By contrast, the use of the device in patients with high ejection fraction and inducible arrhythmia that is controlled by drugs is associated with an incremental cost effectiveness of around £700 000 per year of life saved.

The 548 baseline results used no fewer than 106 different methodological scenarios, and consequently a “reference case” methodological scenario was applied retrospectively to each article; this resulted in a total of 333 methodologically comparable baseline results. These results were converted to a common cost base year and ranked to give a comprehensive “league table” of results for the United Kingdom. Of the 333 results, 61 reported an associated range of high and low values to represent uncertainty. Alternative rankings based on the high or low values from this range showed that there could be considerable disruption to the ranked order based on the baseline point estimates only. This is illustrated by figure 1, which shows the rank ordering of these 61 results by their baseline values and by the highest value from their range. This analysis of UK studies reporting the ranges of sensitivity analyses raises the further concern that the median number of variables included in the sensitivity analysis was just two. Therefore, the ranges of

values shown in figure 1 are likely to be less than if a comprehensive analysis of all uncertain variables had been conducted. Clearly, this would further increase the potential for the rank order to vary depending on the value chosen from the overall range.

Cost data at patient level

Of the 492 studies on the database, only 53 had patient level cost data and just 25 of these reported some measure of cost variance. Eleven reported only ranges, which are of limited usefulness in quantifying variance. Five articles gave a standard error, seven a standard deviation, and only four studies (< 1%) had calculated 95% confidence intervals for cost.

In the five datasets of cost at the patient level, analysis indicated that many cost data were substantially skewed in their distribution. This may cause problems for parametric statistical tests for the equality of two means. One method for dealing with this is to transform the data to an alternative scale of measurement—for example by means of log, square root, or reciprocal transformations. However, our analysis of these data indicated that although a transformation may modestly improve the statistical significance of observed cost differences or may reduce the sample size requirements to detect a specified difference, it is difficult to give the results of a transformed or back transformed scale a meaningful economic interpretation, especially if we intend to use the cost information as part of a cost effectiveness ratio. It would be appropriate to use non-parametric bootstrapping to test whether the sample size of a study’s cost data is sufficient for the central limit theorem to hold, and to base analyses on mean values from untransformed data.

Estimating confidence intervals for cost effectiveness ratios

Finally, our review identified a number of different methods for estimating confidence intervals for cost effectiveness ratios that have appeared in the recent literature,⁹⁻¹⁴ and we applied each of these methods to one of the five datasets listed above.¹⁵ These different methods produced very different intervals. Examination of their statistical properties and evidence from recent Monte Carlo simulation studies^{14 16} suggests that many of these methods may not perform well in some circumstances. The parametric method based on Fieller’s theorem and the non-parametric approach of

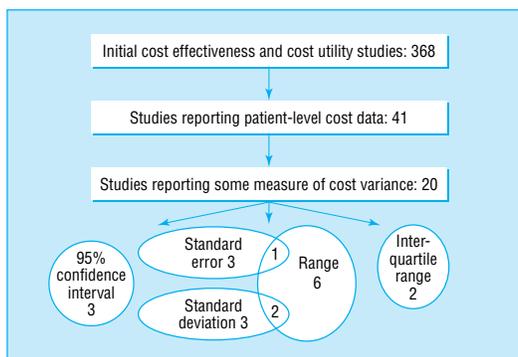


Fig 2 The handling of cost variance by studies reporting patient level cost data

bootstrapping have been shown to produce consistently the best results in terms of the number of times, in repeated sampling, the true population parameter is contained within the interval.^{14 16}

Recommendations

Uncertainty in economic evaluation is often handled inconsistently and unsatisfactorily. Recently published guidelines should improve this situation, but we emphasise the following:

- Ensure that the potential implications of uncertainty for the results are considered in all analyses;
- When reporting cost and cost effectiveness information, make more use of descriptive statistics. Interval estimates should accompany each point estimate presented;
- Sensitivity analyses should be comprehensive in their inclusion of all variables;
- Cost and cost effectiveness data are often skewed. Significance tests may be more powerful on a transformed scale, but confidence interval should be reported on the original scale. Even when data are skewed, economic analyses should be based on means of distributions;
- Where patient level data on both cost and effect are available, the parametric approach based on Fieller's theorem or the non-parametric approach of bootstrapping should be used to estimate a confidence interval for the cost effectiveness ratio;
- When comparing results between studies, ensure that they are representative;
- Using a methodological reference case when presenting results will increase the comparability of results between studies.

This article is adapted from *Health Services Research Methods: A Guide to Best Practice*, edited by Nick Black, John Brazier, Ray Fitzpatrick, and Barnaby Reeves, published by BMJ Books.

Competing interests: None declared.

- 1 Briggs AH. *Handling uncertainty in the results of economic evaluation*. London: Office of Health Economics, 1995. (OHE briefing paper No 32.)
- 2 Manning WG, Fryback DG, Weinstein MC. Reflecting uncertainty in cost-effectiveness analysis. In: Gold MR, Siegel JE, Russell LB, Weinstein MC, eds. *Cost-effectiveness in health and medicine*. New York: Oxford University Press, 1996:247-75.
- 3 Briggs AH, Gray AM. Handling uncertainty when performing economic evaluations of health care interventions: a systematic review with special reference to the variance and distributional form of cost data. *Health Technol Assess* 1999;3(2).
- 4 Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. *BMJ* 1996;313:275-83.
- 5 Gold MR, Siegel JE, Russell LB, Weinstein MC, eds. *Cost-effectiveness in health and medicine*. New York: Oxford University Press, 1996.
- 6 Canadian Coordinating Office for Health Technology Assessment. *Guidelines for the economic evaluation of pharmaceuticals: Canada*. 2nd ed. Ottawa: CCOHTA, 1997.
- 7 Drummond MF, O'Brien B, Stoddart GL, Torrance GW. *Methods for the economic evaluation of health care programmes*. 2nd ed. Oxford: Oxford University Press, 1997.
- 8 Anderson MH, Camm AJ. Implications for present and future applications of the implantable cardioverter-defibrillator resulting from the use of a simple model of cost efficacy. *Br Heart J* 1993;69:83-92.
- 9 O'Brien BJ, Drummond MF, Labelle RJ, Willan A. In search of power and significance: issues in the design and analysis of stochastic cost-effectiveness studies in health care. *Med Care* 1994;32:150-63.
- 10 Wakker P, Klaassen M. Confidence intervals for cost-effectiveness ratios. *Health Econ* 1995;4:373-82.
- 11 Van Hout BA, Al MJ, Gordon GS, Rutten FF. Costs, effects and C/E-ratios alongside a clinical trial. *Health Econ* 1994;3:309-19.
- 12 Chaudhary MA, Stearns SC. Estimating confidence intervals for cost-effectiveness ratios: an example from a randomized trial. *Stat Med* 1996;15:1447-58.
- 13 Briggs AH, Wonderling DE, Mooney CZ. Pulling cost-effectiveness analysis up by its bootstraps: a non-parametric approach to confidence interval estimation. *Health Econ* 1997;6:327-40.
- 14 Polsky D, Glick HA, Wilke R, Schulman K. Confidence intervals for cost-effectiveness ratios: a comparison of four methods. *Health Econ* 1997;6:243-52.
- 15 Fenn P, McGuire A, Phillips V, Backhouse M, Jones D. The analysis of censored treatment cost data in economic evaluation. *Med Care* 1995;33:851-63.
- 16 Briggs AH, Mooney CZ, Wonderling DE. Constructing confidence intervals around cost-effectiveness ratios: an evaluation of parametric and non-parametric methods using Monte Carlo simulation. *Stat Med* (in press).

How the defibrillator saved a patient's life

Initially it was quite a struggle just getting the partners to agree that purchasing a defibrillator would benefit the practice. We did not even have to pay as the Friends of the Health Centre kindly raised the money.

The equipment was installed in the nurses' treatment room and gradually gathered dust. "Does the defibrillator work?" and "I bet the batteries aren't charged" were some of the jocular comments from the partners.

We had a couple of attempts at resuscitation, the equipment worked well, but unfortunately the patient did not survive. It was decided to hold a training day on resuscitation for the nurses. The alarms sounded, I rushed to the treatment room only to find that it was a mock emergency.

In the middle of a busy afternoon surgery the same day the alarm went off again and there was an urgent telephone call. When I arrived several partners and nursing staff were in the middle of full cardiopulmonary resuscitation. The patient had been sent down from the doctor's surgery to the treatment room for an electrocardiogram as he had chest pain and had collapsed. The tracing showed ventricular fibrillation. Bring out the defibrillator! Charge to 200 deliver shock! It's just like *ER!* Unfortunately, the patient was unstable; there were further episodes of ventricular fibrillation and further defibrillation. As a former medical registrar it started to flood back. We need lignocaine, but what is the dose? It was like the blind leading the blind.

Four cardioversions later the ambulance arrived. Was he stable enough to transfer to our local hospital? It was decided that I should accompany the patient in the ambulance; this was just as well as he had two further arrests in the ambulance requiring defibrillation. An emergency stop as a bus pulled out in front of us hurled the patient forward into my lap. But he survived, and as he was only 40 with two children he was eternally grateful.

What have we learnt? Clearly, we need more training in resuscitation. We now have a very persuasive argument for the partner who said that we did not need a defibrillator as the ambulance always carries one. Our Friends of the Health Centre are now saving to buy us a better model that can record the cardiac rhythm through the paddles.

Alexander Williams, *general practitioner, Exeter*

We welcome articles up to 600 words on topics such as *A memorable patient, A paper that changed my practice, My most unfortunate mistake*, or any other piece conveying instruction, pathos, or humour. If possible the article should be supplied on a disk. Permission is needed from the patient or a relative if an identifiable patient is referred to. We also welcome contributions for "Endpieces," consisting of quotations of up to 80 words (but most are considerably shorter) from any source, ancient or modern, which have appealed to the reader.