

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.

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 Desenclos 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.



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