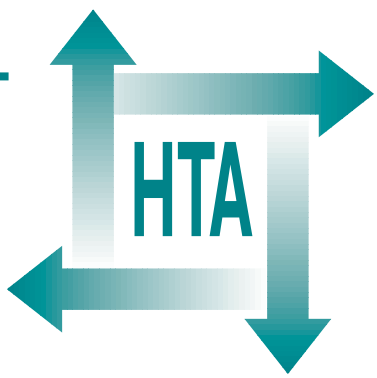
Health Technology Assessment 1999; Vol. 3: No. 5

Review

Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review

OC Ukoumunne
MC Gulliford
S Chinn
JAC Sterne
PGJ Burney

Health Technology Assessment
NHS R&D HTA Programme



Standing Group on Health Technology

Current members

Chair:
Professor Sir Miles Irving,
Professor of Surgery, University of Manchester, Hope Hospital, Salford

Professor Martin Buxton,
Professor of Economics, Brunel University

Professor Francis Creed,
School of Psychiatry & Behavioural Sciences, University of Manchester

Professor Charles Florey,
Department of Epidemiology & Public Health, Ninewells Hospital & Medical School, University of Dundee

Professor John Gabbay,
Director, Wessex Institute for Health Research & Development

Professor Sir John Grimley Evans,
Department of Geriatric Medicine, Radcliffe Infirmary, Oxford

Dr Tony Hope,
The Medical School, University of Oxford

Professor Richard Lilford,
Regional Director, R&D, West Midlands

Dr Jeremy Metters,
Deputy Chief Medical Officer, Department of Health

Professor Maggie Pearson,
Regional Director of R&D, NHS Executive North West

Mr Hugh Ross,
Chief Executive, The United Bristol Healthcare NHS Trust

Professor Trevor Sheldon,
Director, NHS Centre for Reviews & Dissemination, University of York

Professor Mike Smith,
Director, The Research School of Medicine, University of Leeds

Dr John Tripp,
Department of Child Health, Royal Devon & Exeter Healthcare NHS Trust

Professor Tom Walley,
Department of Pharmacological Therapeutics, University of Liverpool

Dr Julie Woodin,
Chief Executive, Nottingham Health Authority

Professor Kent Woods
(**Chair Designate**),
Regional Director of R&D, NHS Executive, Trent

Past members

Dr Sheila Adam,
Department of Health

Professor Angela Coulter,
Director, King's Fund, London

Professor Anthony Culyer,
Deputy Vice-Chancellor, University of York

Dr Peter Doyle,
Executive Director, Zeneca Ltd, ACOST Committee on Medical Research & Health

Professor John Farndon,
Professor of Surgery, University of Bristol

Professor Howard Glennester,
Professor of Social Science & Administration, London School of Economics & Political Science

Mr John H James,
Chief Executive, Kensington, Chelsea & Westminster Health Authority

Professor Michael Maisey,
Professor of Radiological Sciences, Guy's, King's & St Thomas's School of Medicine & Dentistry, London

Mrs Gloria Oates,
Chief Executive, Oldham NHS Trust

Dr George Poste,
Chief Science & Technology Officer, SmithKline Beecham

Professor Michael Rawlins,
Wolfson Unit of Clinical Pharmacology, University of Newcastle-upon-Tyne

Professor Martin Roland,
Professor of General Practice, University of Manchester

Professor Ian Russell,
Department of Health Sciences & Clinical Evaluation, University of York

Dr Charles Swan,
Consultant Gastroenterologist, North Staffordshire Royal Infirmary

Details of the membership of the HTA panels, the NCCHTA Advisory Group and the HTA Commissioning Board are given at the end of this report.

How to obtain copies of this and other HTA Programme reports.
An electronic version of this publication, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (http://www.hta.ac.uk). A fully searchable CD-ROM is also available (see below).

Printed copies of HTA monographs cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our Despatch Agents.

Non-UK purchasers will have to pay a small fee for post and packing. For European countries the cost is £2 per monograph and for the rest of the world £3 per monograph.

You can order HTA monographs from our Despatch Agents:

– fax (with **credit card** or **official purchase order**)
– post (with **credit card** or **official purchase order** or **cheque**)
– phone during office hours (**credit card** only).

Additionally the HTA website allows you **either** to pay securely by credit card **or** to print out your order and then post or fax it.

Contact details are as follows:

HTA Despatch
c/o Direct Mail Works Ltd
4 Oakwood Business Centre
Downley, HAVANT PO9 2NP, UK

Email: orders@hta.ac.uk
Tel: 02392 492 000
Fax: 02392 478 555
Fax from outside the UK: +44 2392 478 555

NHS libraries can subscribe free of charge. Public libraries can subscribe at a very reduced cost of £100 for each volume (normally comprising 30–40 titles). The commercial subscription rate is £300 per volume. Please see our website for details. Subscriptions can only be purchased for the current or forthcoming volume.

Payment methods

Paying by cheque
If you pay by cheque, the cheque must be in **pounds sterling**, made payable to *Direct Mail Works Ltd* and drawn on a bank with a UK address.

Paying by credit card
The following cards are accepted by phone, fax, post or via the website ordering pages: Delta, Eurocard, Mastercard, Solo, Switch and Visa. We advise against sending credit card details in a plain email.

Paying by official purchase order
You can post or fax these, but they must be from public bodies (i.e. NHS or universities) within the UK. We cannot at present accept purchase orders from commercial companies or from outside the UK.

How do I get a copy of *HTA on CD*?

Please use the form on the HTA website (www.hta.ac.uk/htacd.htm). Or contact Direct Mail Works (see contact details above) by email, post, fax or phone. *HTA on CD* is currently free of charge worldwide.

The website also provides information about the HTA Programme and lists the membership of the various committees.

Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review

OC Ukoumunne
MC Gulliford
S Chinn
JAC Sterne
PGJ Burney

Department of Public Health Sciences, Guy's, King's and St Thomas' School of Medicine, King's College London, UK

Published April 1999

This report should be referenced as follows:

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

Health Technology Assessment is indexed in *Index Medicus*/MEDLINE and *Excerpta Medica*/EMBASE. Copies of the Executive Summaries are available from the NCCHTA web site (see overleaf).

NHS R&D HTA Programme

The overall aim of the NHS R&D Health Technology Assessment (HTA) programme is to ensure that high-quality research information on the costs, effectiveness and broader impact of health technologies is produced in the most efficient way for those who use, manage and work in the NHS. Research is undertaken in those areas where the evidence will lead to the greatest benefits to patients, either through improved patient outcomes or the most efficient use of NHS resources.

The Standing Group on Health Technology advises on national priorities for health technology assessment. Six advisory panels assist the Standing Group in identifying and prioritising projects. These priorities are then considered by the HTA Commissioning Board supported by the National Coordinating Centre for HTA (NCCHTA).

This report is one of a series covering acute care, diagnostics and imaging, methodology, pharmaceuticals, population screening, and primary and community care. It was identified as a priority by the Methodology Panel and funded as project number 94/09/01.

The views expressed in this publication are those of the authors and not necessarily those of the Standing Group, the Commissioning Board, the Panel members or the Department of Health. The editors wish to emphasise that funding and publication of this research by the NHS should not be taken as implicit support for the recommendations for policy contained herein. In particular, policy options in the area of screening will be considered by the National Screening Committee. This Committee, chaired by the Chief Medical Officer, will take into account the views expressed here, further available evidence and other relevant considerations.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

The editors have tried to ensure the accuracy of this report but cannot accept responsibility for any errors or omissions. They would like to thank the referees for their constructive comments on the draft document.

ISSN 1366-5278

Published by Core Research, Alton, on behalf of the NCCHTA.
Printed on acid-free paper in the UK by The Basingstoke Press, Basingstoke.

Copies of this report can be obtained from:

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 1703 595 639 Email: hta@soton.ac.uk
http://www.soton.ac.uk/~hta

Contents

List of abbreviations

BMI	body mass index
BRHS	British Regional Heart Study*
CF	correction factor
COMMIT	Community Intervention Trial for Smoking Cessation
DBP	diastolic blood pressure*
df	degree of freedom*
DHA	district health authority*
EM	expectation maximisation*
ERIC	Education Resources Information Centre
FHSA	family health service authority*
FML	full maximum likelihood*
GEE	generalised estimating equation
GP	general practitioner
HC	health centre*
HSI	health service indicators*
HSE	Health Survey for England*
IF	inflation factor
IGLS	iterative generalised least squares*
IGT	impaired glucose tolerance
MeSH	medical subject headings
MML	marginal maximum likelihood*
MQL	marginal quasi-likelihood
PHCDS	Public Health Common Data Set*
PQL	penalised quasi-likelihood
RCP	Royal College of Physicians*
REML	restricted maximum likelihood*
RHA	regional health authority*
SBP	systolic blood pressure*
TCR	Thames Cancer Registry*
TPP	total-purchasing pilot

* Used only in tables

Executive summary

Background

Health technology assessment often requires the evaluation of interventions which are implemented at the level of geographical area or health service organisational unit. Examples include health promotion interventions implemented in schools, workplaces or neighbourhoods, screening programmes in health authority populations, and healthcare interventions in general practices or hospitals. Interventions like these are implemented for clusters of individuals. Evaluation of cluster-based interventions presents a number of difficulties but some evidence suggests these are not always addressed in an optimal manner.

Aims and objectives

This report describes a systematic review of methods for evaluating cluster-based interventions. There were three objectives:

- to review the methodological literature and synthesise the findings into a checklist for practical use
- to evaluate existing practice in healthcare evaluation
- to present intraclass correlations for a range of outcome variables at different levels of organisational clustering in order to provide information for the design of future cluster-based studies.

Methods

- The review focused on methods for evaluating health and healthcare interventions that are implemented for clusters of patients or healthy individuals. References were obtained by handsearching journals, searching electronic databases, screening cited references, contacting expert informants, and searching the world wide web. Synthesis into a methodological checklist was by means of qualitative judgements concerning validity.
- A review of seven health science journals in 1996 yielded 56 papers reporting evaluations of cluster-based interventions. Evaluation against the checklist of methodological recommendations identified the main departures from good practice.
- A database of intraclass correlations was compiled by analysing data from a variety of sources.

Methodological recommendations

The main methodological findings of the review were synthesised into a 12-point checklist for investigators.

(1) **Recognise the cluster as the unit of intervention or allocation.** It is important to distinguish between cluster level and individual level intervention, as failure to do so can result in studies which are inappropriately designed or which give incorrect results.

(2) **Justify the use of the cluster as the unit of intervention or allocation.** For a fixed number of individuals, studies in which clusters are allocated are not as powerful as traditional clinical trials in which individuals are randomised. The decision to allocate at cluster level should be justified on theoretical, practical or economic grounds.

(3) **Include a sufficient number of clusters.** Evaluation of an intervention implemented in a single cluster will not usually give generalisable results. Valid designs should include a control group not receiving the intervention. Both intervention and control groups should include enough clusters to allow the effect of intervention to be distinguished from natural variability among clusters. Studies with fewer than four clusters per group are unlikely to yield statistically significant results, and more clusters will be required if relevant intervention effects are small.

(4) **Randomise clusters wherever possible.** The need for randomisation is generally accepted in the evaluation of individual level interventions but randomisation of clusters has not been practised as often as it should be in the evaluation of cluster-based interventions. Because of the risk of bias, use of quasi-experimental or observational designs should always be justified.

(5) **In non-randomised studies include a control group.** When randomisation is not feasible, a control group should be included. Each

group should include a sufficient number of clusters (see point 3). The clusters allocated to groups should be stratified for important prognostic factors so far as possible (see point 8) and a wide range of confounders should be measured. Outcome variables should be measured before and after the intervention.

(6) In single group studies include repeated measurements over time. Sometimes it is not feasible to include a control group, as, for example, when a new policy is implemented at national level. In this case, repeated assessments should be made both before and after the intervention in order to control for secular changes in the outcome.

(7) Allow for clustering when estimating the required sample size. The total number of individuals required can be estimated by multiplying the result of a standard sample size calculation by the design effect. This will require an estimate of the intraclass correlation coefficient, which should be obtained from previous studies.

(8) Consider the use of pairing or stratification of clusters where appropriate. Cluster-based evaluations often include small numbers of clusters, and simple randomisation is unlikely to yield groups that are balanced with respect to cluster level baseline characteristics. Stratification or pairing of clusters according to characteristics that are associated with the outcome may reduce error in randomised studies and reduce bias in non-randomised studies. Limitations of the paired, or matched, design are underappreciated.

(9) Consider different approaches to repeated assessments in prospective evaluations. Either cohort or repeated cross-sectional designs may be used to sample individuals in studies with follow-up. The cohort design is more applicable to individual level outcomes, and may yield more precise results but is more susceptible to bias. The repeated cross-sectional design is more appropriate when outcomes will be aggregated to cluster level; it is usually less powerful but is less susceptible to bias.

(10) Allow for clustering at the time of analysis. Standard statistical methods applied to individual level outcomes should not be used because they will give confidence intervals that are too narrow and p values that are too small. There are three valid approaches to analysis: cluster level analysis, in which the cluster means or proportions are used as units of analysis; adjusted individual level analysis, in which standard univariate statistical methods are adjusted for the design effect; regression methods for clustered data, which allow for both individual and cluster level variation (hierarchical analysis). When the number of clusters is small, cluster level analysis will be most appropriate because between-cluster variation cannot be estimated with sufficient precision to implement analyses at the individual level. Regression methods for clustered data will usually be required for non-randomised designs.

(11) Allow for confounding at both individual and cluster level. Standard multiple regression methods are not appropriate. Use of regression methods for clustered data will allow the incorporation of both individual and cluster level confounders in the analysis. This approach will increase precision in randomised studies and reduce bias in non-randomised designs.

(12) Include estimates of intraclass correlation and components of variance in published reports. In order to provide information that may be used to estimate sample size requirements for future studies, estimates of the intraclass correlation coefficient should be included in published reports.

Case study: a review of seven health science journals

A review of 56 papers reporting evaluations of cluster-based interventions from seven health science journals showed that the present level of adherence to the methodological recommendations of the review was low. The main departures from recommendations were the evaluation of interventions in small numbers of clusters, and the incorrect use of standard methods for individual level analysis.

A database of intraclass correlation coefficients

In order to provide information which may be used in the design of future studies, the report presents intraclass correlation coefficients and components of variance for a range of outcomes in five areas: cardiovascular and lifestyle, cancer, respiratory, health service activity, and other. For community-based studies, data are presented for individuals clustered at the level of household, postcode sector and district and regional health authority. For healthcare-based studies, data are presented for clustering at the level of general practice, hospital, district health authority and family health services authority.

Chapter 1
Introduction

Areas and organisations as clusters of individuals

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

Rationale for cluster-based studies

Traditional clinical epidemiological approaches to healthcare evaluation have regarded the individual subject as the unit of intervention and analysis, as, for example, in a clinical trial to evaluate the efficacy of drug treatment. In area-wide and organisation-based evaluations (cluster-based evaluations) the unit of intervention is a geographical or organisational cluster. There are both theoretical and practical reasons why intervention and evaluation at area or organisation level may be appropriate.

Individuals do not exist in isolation. Changes in health policy or health service organisation and practice are usually implemented within areas and health organisations and not at individual level. Evaluation at area or organisation level is appropriate from a theoretical perspective.[1–5]

From a practical perspective, there are several reasons why cluster-based evaluation may be appropriate. An intervention may necessarily affect all members of a geographical area.[6] Examples are a regional trauma centre directed at the catchment population of a region or use of an advertising campaign to reduce smoking prevalence. Even when individual allocation is feasible, there may be ethical problems associated with treating some subjects within clusters differently to others.[1–5,7–13] In some instances, it may not be convenient from an administrative or political viewpoint to allocate members of the same organisation to different intervention groups.[2,6,9,11] For example, in a study to evaluate a dietary intervention for schoolchildren, it will usually be easier to allocate clusters, such as classes or schools, rather than allocate pupils within the same school to different dietary choices. In studies of general practice, GPs may be unwilling to randomise individual patients, so randomising GPs themselves becomes more feasible for political and practical reasons.[14] Cluster-based evaluation is also more convenient if there is no sampling frame from which individuals may be selected.

Cluster-based evaluations provide opportunities to include cluster level outcomes and cluster level confounding variables,[15,16] this is not usually possible in studies with individual level allocation.

Use of cluster-based evaluation may reduce contamination between intervention groups.[4,7,13,17] Contamination can be a problem if individuals within the same community are randomised to different groups, particularly when blinding is impossible. Conversely, 'contamination within clusters' is beneficial in the context of community-wide evaluations.[2,4] Subjects in the same cluster tend to mix, and as a consequence the intervention is diffused more efficiently. The success of an intervention may rest upon the ability to change the behaviour of the cluster overall.

Cluster-based evaluations can be more cost-effective than trials in which the intervention has been applied to individuals.[1–3,5–7,9,16,18–21]

Problems of cluster-based studies

Cluster-based evaluations are often appropriate for theoretical, practical or financial reasons, but this type of evaluation presents special problems which need to be addressed before a study can produce valid, generalisable results.

There are three main methodological problems associated with cluster-based evaluation:

- the level of intervention may differ from the level of evaluation
- there may be a small number of units of intervention (clusters)

- outcomes of individuals are often correlated within clusters.

Distinction between cluster level and individual level evaluation

A basic problem in cluster-based evaluation is that the level of intervention is often different from the level of measurement (*Table 1*). Donner and Klar[11] observed that cluster level interventions present a less serious problem to investigators when inferences from evaluation are intended at cluster level, because each cluster can then be considered as the individual unit of observation or analysis. For example, the consequences of different methods of funding primary care services might be evaluated according to whether certain types of practice had better practice facilities. Cluster level interventions are often designed to modify individual level outcomes. For example, a study might aim to determine whether setting up a general practice diabetic clinic resulted in better blood glucose control in the practice's patients. In this instance intervention is at cluster level but inferences are intended at individual level.[11] This leads to the 'unit of analysis problem', where standard individual level analyses are performed inappropriately in a cluster-based intervention study. Whiting-O'Keefe and colleagues found that this type of error was present in 20 out of 28 healthcare evaluations which they studied.[22]

Designing studies with small numbers of clusters

Area-wide and organisation-based evaluations typically include only small numbers of clusters. For example, the North Karelia Project[5] included one intervention and one control region, the Stanford project included five cities,[24] and a recent evaluation of a regional trauma centre included one intervention and two control regions.[25] The total number of individuals included in each of these studies was large. However, when cluster level interventions are being evaluated, the power of the study is determined more by the number of clusters than by the number of individuals in the study. One of the challenges presented by cluster-based evaluations is to determine how studies may be designed most efficiently using small numbers of clusters.

Correlation of outcomes within clusters

Standard statistical methods are based on the assumption that individual responses are independent of one another. In the context of organisation level evaluations, responses are rarely independent because areas and organisations tend to contain individuals who are more similar to each other than they are to individuals in other areas or organisations. In other words, because individual responses usually show some correlation within clusters, variation between clusters is greater than variation within clusters.[11,12,22] There are at least three reasons why this might be the case.[12,26–28] Firstly, subjects may select the cluster to which they belong. For example, a patient's choice of general practice may be associated with characteristics such as age, gender or ethnic group. Secondly, cluster level variables might influence all cluster members in a similar direction and each cluster will be subject to distinct influences. For example, patient outcomes may differ systematically among surgeons. Thirdly, individuals within clusters may interact, influence each other and thus tend to conform, as, for example, when individuals within a school class respond to a health promotion message.

The degree of correlation between individuals depends on the type of cluster and the nature of the outcome variable. In general, because members of the same cluster are unlikely to be independent with respect to health outcomes,[29] standard statistical methods for calculating sample sizes, assessing

TABLE 1 *Comparison of levels of intervention and levels of evaluation*

Level of evaluation	Level of intervention	
	Individual	**Area or organisation**
Individual	Example: Does treatment with β interferon decrease morbidity from mutiple sclerosis?	Examples: Does setting up a nurse run asthma clinic improve patient health outcomes? Does providing a baby-friendly environment in hospital increase mothers' success at breast-feeding?
Area or organisation		Examples: Do smoking control policies increase the proportion of smoke free work places? Do fundholding general practices develop better practice facilities than non-fundholders?

Source: adapted from McKinlay[23]

power and analysing data are not appropriate for use in cluster-based evaluations. Because the variance of the outcome is inflated by between-cluster variation, use of standard methods will result in the required sample size being under-estimated, and the significance of the intervention will be overestimated. In other words, standard sample size calculations will result in studies with insufficient power to detect an intervention effect and standard statistical tests will tend to reject the null hypothesis too often.

Each individual in a cluster randomised study contributes less information than for a study in which the subjects themselves are randomised. For this reason the decision to use clusters as the unit of intervention should always be explicitly justified in study proposals and reports.[9,12,13]

Aims and objectives

The aim of the project was to conduct a systematic review of methods for evaluation of area-wide and organisation-based interventions.

Specific objectives

The specific objectives of the project were to:

- produce a systematic methodological review, synthesised into a set of methodological guidelines
- present an evaluation of existing practice in healthcare evaluation so as to identify the main departures from good practice
- present estimates of between-cluster variation for a range of outcome variables at different levels of organisational clustering.

Chapter 2
Methods

The systematic review has been shown to be an important tool for analysing and presenting evidence of the effectiveness of interventions by pooling results from randomised trials, and for establishing associations of risk factors with disease outcomes by pooling the results of observational studies. Systematic reviews have only recently come to be used in the analysis of methodological problems, and there are presently no standard guidelines for the conduct of methodological reviews. Scientific strategies must still be applied, however, to avoid error and bias.[30]

Three characteristics distinguish the approach to methodological systematic reviews:

- Methodological reviews have a broad focus, and it is important to chart the subject area of the review in order to ensure that important aspects of the problem are not neglected.
- Because the focus for the review is broad, a systematic search will usually generate a large number of citations of varying relevance and degree of overlap. Some selection must be made so that the final review is based on a smaller number of key references.
- Finally, qualitative methods of synthesis are required, so that methodological information can be assembled into a narrative review which includes a succinct set of recommendations.

Hutton and Ashcroft recently discussed how systematic reviews can best be carried out.[31] They pointed out that a fully comprehensive review is rarely feasible, and they recommended that reviewers should use approaches that are relevant to the purpose of the review. The main purpose of our review was to provide guidance for optimal practice in the design, conduct and analysis of evaluation of healthcare interventions at area and organisation level. We emphasised relevant methodological developments. There are areas with some relevance to the focus of the review, for example the design and analysis of complex surveys, which were not reviewed in detail. This was because we judged that these were not areas which needed to be evaluated in depth in order to meet the purposes of the review. Similarly, we judged that it was neither necessary nor feasible to attempt a fully systematic review of methodological literature in multilevel modelling.

Definition of focus for the review

The review focused on methods for evaluating health or healthcare interventions that are implemented for clusters of patients or healthy individuals. We aimed to include methods appropriate for both experimental studies and observational evaluations of existing services. We did not aim to investigate qualitative methods nor methods of health economic evaluation.

We charted three main areas for the review. Under study design we considered issues relating to randomisation, design of non-randomised studies, sample size and power calculations, stratification and matching, and methods of sampling individuals within clusters. Under analysis we considered methods of univariate analysis at the individual and cluster level, hypothesis testing and estimation, approaches for stratified or paired designs, and methods for controlling for confounding variables at the individual and cluster levels. We considered each of these issues for different types of data, including continuous, binary, ordinal, and time to event data. Finally we considered a number of issues relating to the conduct of cluster-based evaluations.

Search strategy

The search was restricted to English language papers on the grounds that including non-English language papers greatly increases the difficulty and cost of a review.[32]

Initial searches were biased towards papers published in journals, but all other sources of knowledge were considered for inclusion in the

review, including books, newsletters, conference papers and personal communications from key informants.

Multiple methods of ascertainment were used to identify relevant published work. The main search methods used were handsearching, computer-assisted searching, collection of relevant cited work and papers recommended by key informants.

Personal collections

Initially we had available the authors' personal collections of references. Professor Allan Donner also provided a personal collection of references.

Handsearches

Initial handsearches were made of the journal *Statistics in Medicine* for the period January, 1992 to July, 1996. A subsequent search was made of the following journals for the period July 1996 to July 1997; *Statistics in Medicine, American Journal of Epidemiology, International Journal of Epidemiology, Journal of Clinical Epidemiology, American Journal of Public Health, Journal of Epidemiology and Community Health* and the *Journal of Community Health*. At the time of revision further searches of selected journals were continued up to November 1998.

Electronic databases

The **MEDLINE, EMBASE** ***(Excerpta Medica)*** **and ERIC (Education Resources Information Centre)** databases were searched. While the first two are biomedical databases, the ERIC database contains material relevant to educational research. The latter was included because early advances in modelling data with a multilevel structure were made in this field.

MEDLINE

We formulated a search strategy using papers which were available from our personal collection and from the handsearch. We identified those papers that were included in MEDLINE for the years 1992–1996, and from the MEDLINE abstracts we identified the most frequently used key words (medical subject headings (MeSH)). They were found to be **cluster analysis**, **randomised controlled trials**, **research design** and **statistical models**. We combined these with selected textwords and applied the resulting search strategy to all available years on MEDLINE from 1966 onwards. We specifically searched for papers describing non-randomised studies using '**quasi-experiment**' and '**non-randomised**' as text terms. The strategy used is shown in *Box 1*.

BOX 1 MEDLINE search strategy

(1) Cluster analysis (MeSH)
(2) Randomised controlled trials (MeSH)
(3) Research design (MeSH)
(4) Statistical models (MeSH)
(5) – (1) or (2) or (3) or (4)
(6) Cluster (textword)
(7) Community (textword)
(8) Clusters (textword)
(9) Clustered (textword)
(10) Clustering (textword)
(11) Community (textword)
(12) – (6) or (7) or (8) or (9) or (10) or (11)
(13) – (5) and (12)

ERIC

The ERIC database, accessed via Brunel Learning and Information Services, was used to capture papers on random effects modelling or multilevel modelling as it is also known. Four general terms were identified as potentially useful search terms and formed the cornerstone of the search strategy: multilevel model; random coefficient model; hierarchical model; and hierarchical linear model. Searches made in the ERIC database were for all papers containing the specified terms in the title, abstract, notes or descriptors. The full list of searches implemented is given in *Box 2*. Papers were selected for the study upon inspection of the titles and abstracts. The ERIC database contains papers from 1966 onwards.

BOX 2 ERIC searches

(1) Multilevel model(s) or model(l)ing
(2) Random coefficient model(s) or model(l)ing
(3) Hierarchical linear model(s) or model(l)ing
(4) Hierarchical model(s) or model(ling)
(5) Multilevel, research design
(6) Random coefficient, research design
(7) Hierarchical, research design
(8) Multilevel, sample size
(9) Multilevel, design
(10) Random coefficient, sample size
(11) Random coefficient, design
(12) Hierarchical, sample size
(13) Hierarchical, design

BIDS: Excerpta Medica *(EMBASE), Science Citation and Social Science Citation Index*

The papers in the database at the time of the BIDS search were used to identify key textwords for searches. The search implemented in BIDS (1980 onwards) was such that all papers containing the search textword in the title or abstract were retrieved along with those that were classified under the same **minor** or **major keyword**. Over 100 separate search terms were used, and for this reason they are not all listed here. We specifically searched the Science Citation Index, the Social Science Citation Index and EMBASE for papers on **quasi-experimental** and **non-randomised** studies.

Internet searches

Some relevant materials were identified through the world wide web. *Ad hoc* searches were carried out with the objective of identifying appropriate statistical software for cluster-based studies, mostly through the web site of the Multilevel Models Project (http://www.ioe.ac.uk/multilevel/).

Criteria for retrieval, validation and synthesis

The search process identified a very large number of potential papers for review. We initially inspected the titles and abstracts to evaluate their relevance to the focus of the review. We went on to retrieve relevant papers, each of which was reviewed for validity. The primary reviewer was OU, secondary reviewers were MG, SC and JS. Papers were assessed against conventional epidemiological and statistical principles, and qualitative judgements were made of their validity. A narrative review was drafted in which the methods proposed in the most relevant and valid papers were recommended for adoption. Our general approach could be classified as one of 'best evidence synthesis'.[33]

Chapter 3

Study design

Healthcare interventions and experimental interventions

Designs used in health services research are usually classified using terminology borrowed from analytical epidemiology, but this terminology cannot be used without ambiguity. Healthcare interventions may be evaluated either in the context of experimental studies, sometimes called intervention studies, or in the context of observational designs. The former should be used to evaluate new innovations while the latter are more appropriately used to evaluate the quality of existing services. In its epidemiological sense the term 'observational' refers to studies in which there is no intervention. In order to avoid confusion, it is important to make the distinction between **healthcare interventions** and **experimental interventions**. Observational studies of healthcare interventions do not incorporate experimental interventions initiated by the investigator. A conventional classification of epidemiological study designs is shown in *Figure 1*.

For both experimental and observational designs we are concerned with studies in which the unit of intervention or observation is a cluster of individuals, such as a geographical area or unit of health service organisation. For intervention studies, **clusters** of **individuals** are included in **groups** which receive the same **intervention**.

Minimum number of clusters

Health services researchers are often asked to evaluate interventions implemented in single clusters. Typical examples include the evaluation of a regional trauma centre[25] or The North Karelia Project.[5] This type of evaluation is sometimes strengthened by including more than one control cluster.[25] A few commentators have suggested that a one-to-one comparison of clusters can be rigorous enough to generalise the results to a whole country or larger area so long as the clusters are typical,[3,5] but this type of study suffers from serious limitations.

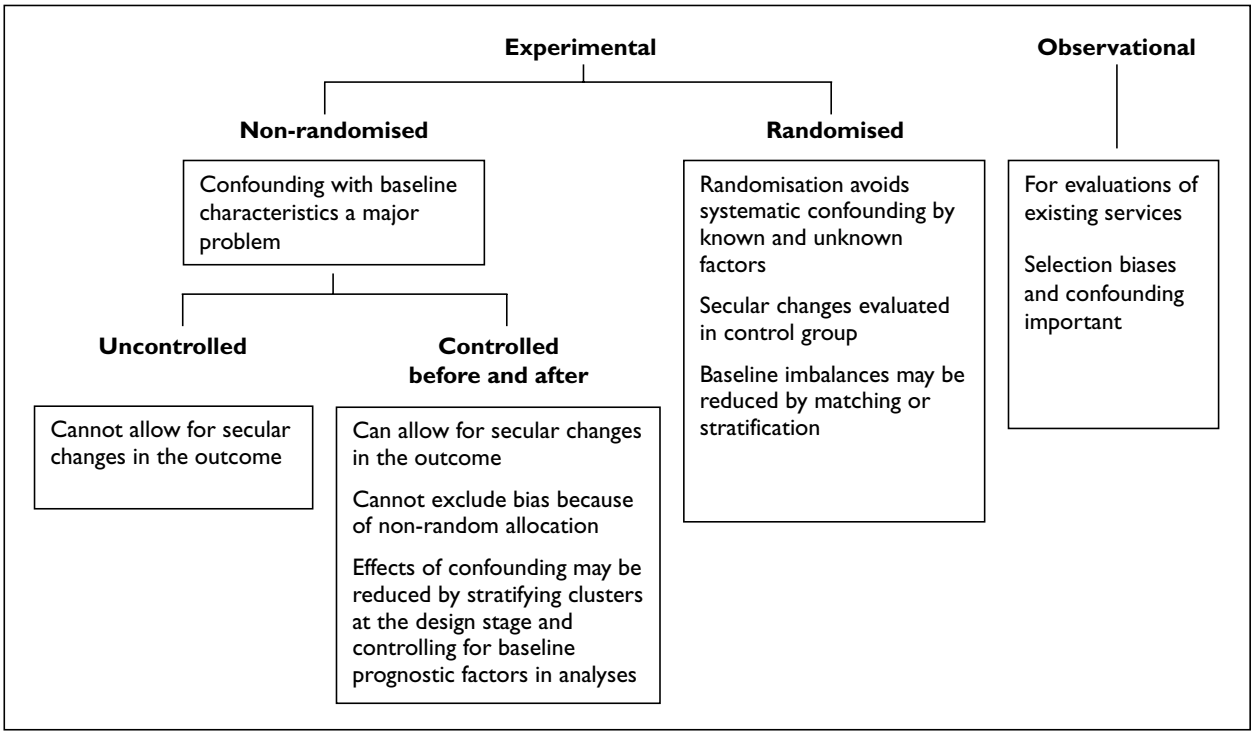


FIGURE 1 *A classification of study designs*

The implementation of an intervention in one or a few clusters usually lacks generalisability and provides a weak basis on which to generate findings for policy formulation. If only one cluster is allocated to each group the intervention effect cannot be separated from the natural variability among clusters.[34] At least three or four clusters per group will be required to provide sufficient degrees of freedom to detect an intervention effect using standard statistical methods applied at the cluster level. Studies with as few as six[35] clusters per group have been used to demonstrate the effects of an intervention, but more clusters will often be needed, particularly when small intervention effects are relevant. Even as many as ten clusters per group may not be enough to estimate the between-cluster variation with sufficient precision to allow analysis at the individual level.

Generally a minimum of three or four clusters should be allocated to each group if cluster level analyses are intended, and ideally many more if individual level analyses are intended. Researchers need to advise health decision makers and service providers that, although of local interest, evaluation of interventions implemented in one or two clusters is unlikely to yield decisive, generalisable results.

Example of an intervention implemented in one cluster with a single control cluster: The North Karelia Project[36]

The North Karelia Project was a community-based cardiovascular disease control programme which included one intervention and one reference community in Finland. Each community was a province with approximately 200,000 inhabitants.

Example of an intervention implemented in one cluster with two control clusters: effectiveness of a regional trauma system in reducing mortality from major trauma[25]

This study evaluated the effect of an experimental regional trauma centre on the survival of patients with major trauma. The trauma centre and five associated district hospitals in the West Midlands were compared with two control regions in Lancashire and Humberside.

Study design and validity

The **validity of a study** may be defined as the degree to which the inferences drawn from it are correct.[37] Two types of validity are recognised in relation to study design, **internal validity** and **external validity**. Internal validity is the extent to which the estimate of intervention effect from a study is unbiased. External validity is the extent to which an estimate of intervention effect from a study may be generalised to a given target population.

The selection of clusters for inclusion in the study has an important bearing on the external validity of a study. A study is more likely to give generalisable results if it can be shown that the units selected for study, and agreeing to participate, were representative of the population of units.[38] Similarly, the intervention should be typical of the one available to the target population. The allocation of clusters to groups has an important bearing on the internal validity of the study. When clusters have been randomly allocated to intervention and control groups and the study has been conducted carefully,[39] the major sources of bias should be avoided. Non-randomised study designs on the other hand are inherently susceptible to several biases.[39,40]

The major threats to the internal validity of a study are:[41]

- '**history**': external events occurring between the pre- and postintervention measurements may influence the outcome in addition to the intervention
- '**maturation**': the passage of time may bring about changes in the experimental units independent of the intervention
- '**testing**': administering a baseline measurement may alter the response to subsequent measurements
- '**instrumentation**': measures used to evaluate outcomes may change over time, for example because disease definitions change or because observers or measuring instruments change
- '**regression to the mean**': experimental units selected on the basis of their extreme scores will tend to give subsequent scores closer to the average
- '**selection bias**': occurs when different types of experimental units are recruited into different groups
- '**differential attrition**': occurs when loss of experimental units from groups is related to the intervention
- '**selection maturation interaction**': occurs when time-dependent changes vary systematically in different types of experimental units.

The potential influence of each source of bias should be considered carefully when appraising studies which aim to evaluate the effectiveness of an intervention. It is important to be aware that some

study designs do not allow for control of particular forms of bias (*Table 2*). For example, studies with a single group will not usually permit control of the influence of external events other than the intervention ('history').

Experimental (intervention) studies: randomised and non-randomised

In general, a valid experimental study design will require two groups (intervention and control), with measurements before and after the intervention, to ensure that each source of bias may be controlled (see *Figure 1*). Randomisation is the only method of allocation that will control for unknown confounders and is to be strongly advocated. When non-randomised study designs are used, careful consideration should be given to the extent to which each form of bias might influence the assessment of intervention effect[41,42] (*Table 2*).

In clinical or laboratory research, the subject usually participates in a study at the invitation of the researcher, and randomised studies may be readily carried out. In 'field' settings, in which many area- or organisation-based evaluations are performed, the evaluator often participates at the invitation of the commissioners or providers of services, and it may be difficult to achieve randomisation in this context.[42] For this reason, some attention is paid here to non-randomised designs. Methods which were developed for sample size estimation and analysis in cluster randomised studies may also be adapted for application to non-randomised designs.

Non-randomised designs

In two monographs, Cook and Campbell[42] and Campbell and Stanley[41] described several non-randomised designs and outlined the biases to which they are susceptible. They made a distinction between pre-experimental and quasi-experimental designs,[41] arguing that pre-experimental designs are unlikely to provide valid evidence for the effectiveness of an intervention, while quasi-experimental designs may be sufficiently resistant to bias to provide valid evidence under some circumstances.

Pre-experimental designs

One-group post-test-only design

In this design, observations are made after an intervention. Because observations are not made before the intervention, it is not certain that the intervention has resulted in change. Because there is no control group, allowance cannot be made for secular changes resulting from maturation, or from the effects of external factors other than the intervention. This design has very limited application in quantitative research, but interpretable information may sometimes be obtained from detailed qualitative case studies.[43]

Example of one-group post-test design: evaluation of total purchasing pilots in England and Scotland.[43] This study evaluated 52 first-wave total-purchasing pilot (TPP) schemes. Evaluation was by means of interviews with key informants and analysis of hospital episode statistics. There was no control group of practices not in TPP schemes, and no data were obtained from participating practices before entry to the TPP schemes. The

TABLE 2 *Classification of non-randomised study designs according to number of groups and timing of observations, and major sources of bias*

Number of groups	Observations	Potential major sources of bias
Intervention group only	After	Selection and attrition of sample; maturation effects; external influences on outcome
	Before and after	Maturation effects; external influences on outcome; effects of testing; secular change in outcome assessment
	Interrupted time series	External influences on outcome; secular change in outcome assessment
Intervention group and non-equivalent control group	After	Maturation effects; selection and attrition of sample
	Before and after	Residual selection bias; interaction of selection and maturation; regression to mean
	Interrupted time series	Residual selection bias; interaction of selection and maturation; regression to mean

Source: adapted from Campbell and Stanley[41]

study presented an analysis of the reported achievements of TPP schemes.

Post-test-only design with non-equivalent control group

In this design, observations are taken after the intervention, in an intervention group and a non-equivalent control group. Because baseline observations are not taken, differences in the outcome between groups may be interpreted either in terms of selection bias or the effect of intervention. Results obtained using this design are usually uninterpretable,[42] and the design is not considered further here.

One-group, pretest post-test design

Observations are made in one group before and after an intervention. There are several reasons why it may not be justified to presume that changes in the outcome are the result of the intervention. There may be secular changes in the outcome caused by maturation of the subjects or by other external factors; regression to the mean may sometimes be important; previous experience of the test might influence subsequent scores; and definitions of the outcome measure may change over time. A control group should be included in order to limit these potential biases.[42]

Example of one-group pretest post-test design: changes in population cholesterol concentrations and other cardiovascular risk factor levels after 5 years of the non-communicable disease intervention programme in Mauritius.[44] A population-wide intervention programme promoting a healthy lifestyle was implemented in the island of Mauritius. Population surveys were carried out before the intervention in 1987 and after in 1992. Comparison of the prevalence of hypertension, cigarette smoking, alcohol consumption, physical activity and serum cholesterol concentrations from the two surveys showed favourable changes. It was concluded that lifestyle intervention projects can have positive effects. This interpretation was plausible but the study suffered from the obvious weakness that there was no information available concerning what would have happened without the intervention.

Quasi-experimental designs

Quasi-experiments evaluate the effect of an **intervention** on the **outcome** for different groups but do not use randomisation to create the comparisons from which the effects of intervention are inferred. The basic problem in interpreting quasi-experimental studies is to separate the effects of intervention from the initial non-comparability of the groups.[42] Two basic types of quasi-experimental design can be distinguished, **non-equivalent group** designs and **interrupted time series** designs.

Non-equivalent group designs

Cook and Campbell[42] used the term **non-equivalent group design** to describe studies in which there are two or more experimental groups which are not formed through randomisation. The simplest form of non-equivalent group design includes intervention and control groups, with observations made before and after the intervention in both groups. By including an untreated control group, this design avoids some of the limitations of the one-group pretest, post-test design. Secular changes resulting from maturation and external factors may be estimated, as may the effects of regression to the mean, repeated testing and changes in methods of outcome measurement. The main threat to validity comes from bias in selection to the intervention and control groups. Post-test differences in outcome may be adjusted for baseline measures by using analysis of covariance. However, because measurements will usually be made with error, adjustment is likely to be incomplete. In addition, the extent of the secular trend may be dependent on baseline factors. In other words there may be an interaction between selection factors and the secular trend. Finally, there may be unknown or unmeasured confounders which may bias the outcome comparison.

Example of non-equivalent group design: effects of the Heartbeat Wales programme over 5 years on behavioural risks for cardiovascular disease.[45] In this study, a health promotion programme was implemented in Wales, starting in 1985, with the aim of improving modifiable behavioural risks for cardiovascular disease. A comparison was made with a reference area (North East England) which was chosen because it was the part of the UK most similar in sociodemographic and health profile to Wales at the 1981 census. Data were collected by means of cross-sectional surveys carried out in the intervention and reference areas in 1985 and 1990. Analysis was by comparison of 15 self-reported indicators of dietary choice, smoking, frequency of exercise and weight between areas. Community level analysis was achieved by comparing nine district health authority areas in Wales with the four counties in the reference area. There were positive changes in outcomes in both areas but no net intervention effect comparing the two areas. Contamination, that is, the possibility that publicity in Wales affected the North East, is a possible explanation for these findings. Note that the effect of intervention is completely confounded by

underlying differences between Wales and the North East England.

Interrupted time series designs

In the interrupted time series design, multiple observations are made before and after an intervention is implemented.[42] The comparisons which are used to gauge the effect of intervention are made within a single group, and the 'control' observations are those made before the intervention is implemented. This design allows secular changes resulting from maturation or regression to the mean to be estimated before the intervention is implemented. The use of repeated measurements before and after the intervention should also allow testing and instrumentation effects to be controlled. As a general rule, studies with three or fewer observations both before and after the intervention are unlikely to give conclusive results.

This design does not allow for the effect of 'history', but if a non-equivalent control group is also included then this design will also allow for more rigorous control of external influences on the outcome. The design can be further elaborated to include removal of the intervention or repeated application of the intervention. Obtaining repeated measurements may be costly if primary research is used, but if routinely collected data are used, then the problems of incompleteness and changing definitions over time may complicate interpretation.

Example of a single group interrupted time series design: the benefit of seat belt legislation in the UK.[46] This study evaluated the effectiveness of legislation enforcing the wearing of seatbelts at reducing road accident fatalities. Legislation was introduced in early 1983. Annual mortality data for road accident deaths in England and Wales were analysed for 7 years before and 5 years after the introduction of the legislation (i.e. from 1977 to 1987). No obvious effect of the intervention was apparent.

Example of an interrupted time series design with several groups: family credit and uptake of school meals in primary school.[47] In 1988, families receiving Family Credit as a welfare benefit in England lost their right to receive free school meals. This study examined the effect of this policy change on the uptake of school meals by children from families receiving Family Credit in comparison with children from families receiving Income Supplement, and those receiving no benefits. Data from annual cross-sectional surveys were analysed for the years 1982 to 1993. The change in legislation resulted in an immediate drop in uptake of school meals by Family Credit children of about 30%. Uptake of school meals in Income Supplement children and children not receiving benefits did not show a marked change.

Limitations of quasi-experimental studies

In quasi-experimental studies, differences in either the intervention effect or the secular trend may result from confounding by differences between the groups. Even after controlling for confounding at the time of analysis, residual confounding and unknown confounders may bias assessment of the effects of intervention. For this reason, randomised designs are to be preferred in health technology assessment, and convincing justification should always be given for the use of non-randomised designs. If randomisation is not feasible, then the effects of confounding can be reduced by using restricted forms of allocation, that is by matching or stratification of clusters which are allocated to the intervention and control groups (see later).

Randomised designs

The purpose of randomisation is to ensure preintervention comparability of study groups. When units are randomly allocated to groups, the groups will be similar on average, and it can be inferred that differences in the outcome are likely to be caused by the intervention. In itself randomisation is not sufficient to ensure that estimated intervention effects are measured without error or bias. This will also depend on other features of the design, conduct and analysis of the study[39] which are included in checklists for the reporting of intervention studies.[38]

Under the completely randomised design (or **unrestricted randomisation**) clusters from a single pool are allocated to groups.[10] Unrestricted allocation is appropriate when there are many clusters available to be allocated.[10,11] A weakness of the approach is that when there are few clusters, treatment groups are likely to be unbalanced with respect to baseline characteristics.[11] **Restricted allocation** may then be used to reduce the extent of baseline imbalances.[10]

Example of cluster randomised study with unrestricted randomisation: effects of diet and exercise in preventing NIDDM in people with impaired glucose tolerance – the Da Qing Impaired Glucose Tolerance and Diabetes Study.[48] In this study among 110,660 men attending 33 health centres, 577 were identified as having impaired glucose tolerance (IGT). Subjects with IGT were

randomised by clinic either to a control group or one of three intervention groups: diet only, exercise only, or diet and exercise. Subjects were followed up at 2 yearly intervals for 6 years to determine whether they developed non-insulin-dependent diabetes. Analysis by clinic showed that each of the intervention groups differed significantly from the control clinics.

Restricted allocation

Under restricted allocation, clusters are first divided into strata according to prognostic characteristics. For example, electoral wards might be stratified according to indicators of social deprivation. Clusters are then allocated to groups within strata. The paired design is a special case of the stratified design in which there are only two clusters per stratum and one cluster is allocated to each group. In paired designs the clusters are individually matched, while in stratified designs they are matched for broader categories.[49] The paired design is often referred to as the matched-pairs design.

Stratification of cluster units in randomised studies is designed to give intervention groups that are more evenly balanced with respect to prognostic variables. This is particularly important in cluster-based studies, which typically include small numbers of clusters, and baseline imbalances may result by chance. Restricted allocation has the effect of increasing statistical power by making estimates of the outcome more precise.[11,12,15,50–52] Restricted allocation can also be used to reduce bias in non-randomised studies.[51,52] Murray stated that non-random assignment of a small number of poorly matched or unmatched units to each group is inadvisable and should be avoided.[53] The Minnesota Heart Health Program provides an example of a quasi-experimental pair-matched study.[54]

Clusters should only be stratified on variables that are highly correlated with the outcome of interest.[10,34,50,51,55] Baseline values for the outcome variable will often be a useful stratifying factor if data are available.[8] In community intervention studies, it is fairly common to match on variables for which routinely collected data are available such as geographical proximity, cluster size, urbanisation or socioeconomic indicators. Unfortunately it may be difficult to anticipate which variables will be related to the outcome of interest. If the stratifying variable is not associated with the outcome, then the potential benefits of restricted allocation will not be realised and stratification will add unnecessary complexity to study design and analysis.

Paired designs

Under the matched-pairs design, clusters are individually matched in pairs with respect to baseline characteristics that are associated with the outcome, and one cluster from every pair is allocated to each group. The Community Intervention Trial for Smoking Cessation (COMMIT)[56] and the British Family Heart Study[57] used matched-pairs designs.

Several methods can be used to identify matched pairs. The first is to divide the population of clusters into relevant strata and then select cluster pairs from each stratum. The second is to identify pairs from clusters that have already been selected for study from the population of interest, and the third is to identify and randomly select relevant pairs. In community intervention studies, the clusters in each pair are not required to be geographically close to each other; indeed, the pairing of clusters that are geographically remote may reduce contamination between the two clusters in each stratum.

The matched-pairs design has important limitations which should be recognised. It may be difficult to find matching variables that can be used to create distinct pairs, particularly for a large number of clusters.[51,55] Another potential problem is that in the event that a particular cluster drops out of the study, the other cluster and thus the entire stratum must be eliminated, causing a reduction in study power.

Matching on variables that are related to the outcome has the effect of reducing the variance of the estimate of intervention effect and increasing the power of the study. However, a loss of degrees of freedom is incurred by using a matched analysis instead of an unmatched one. This loss becomes increasingly important for smaller studies and any gain in power may be cancelled out. The power of a study will inevitably be reduced by matching on a variable that is unrelated to the outcome if a matched analysis is used. For matching to be successful, the decrease in experimental error needs to be sufficiently large to offset the reduction in degrees of freedom. Freedman and colleagues[50] presented a method for estimating the gain in efficiency resulting from matching compared with not matching. They also devised a method for assessing the extent to which matching on the baseline values for the outcome further improves efficiency. A limitation of their method is that it does not consider reduction in power of the matched design due to the loss of degrees of freedom in small sample studies, and therefore it

might lead to the erroneous conclusion that even when there is a relatively small number of clusters, matching results in greater efficiency.[55]

Martin and colleagues,[55] basing their work on the *t* tests for paired and unpaired samples, showed that if the number of study clusters is small then an unmatched randomised design will usually be more efficient because of the loss of degrees of freedom associated with the matched design. Only the strongest of correlates would make matching worthwhile for very small studies. They also showed that for studies with less than around ten pairs of clusters a matched design will often be appreciably less powerful than an unmatched design.[55] Diehr and colleagues[34] obtained much the same result which they generalised to studies in which there are between three and around ten pairs. Diehr and colleagues[34] suggested that in studies with small numbers of clusters, if pre-intervention matching seems to be required at the outset, an unmatched analysis will usually be more powerful because of the increased degrees of freedom compared with the matched analysis.

Klar and Donner[51] cautioned that the matched-pairs design may be less suitable for cluster-based evaluations because of the difficulty of separating the effect of intervention from between-cluster variation for adjusted individual level analysis. This is because between-cluster variation is confounded with the intervention effect, with differences between cluster pairs as well as with between-stratum variation in the intervention effect (i.e. the stratum–intervention interaction).[11,51,58–60] In other words, between-cluster variation cannot usually be estimated within pairs because each cluster within a pair receives a different intervention and between-cluster, within-group variation cannot be estimated due to confounding with differences between pairs. Given that there is a non-zero intervention effect and real differences between strata, an unbiased estimate of between-cluster variation can only be obtained if there is no interaction between intervention effect and stratum. Thus, the consequence of calculating the between-cluster variation from a matched-pairs study is that the variation will usually be overestimated, leading to conservative results in adjusted individual level analyses and overestimates of the required sample size for future studies. Thompson and colleagues[61] recently proposed that random effects meta-analysis can be applied to the analysis of such designs. One variance component can be used to allow for both natural variation between clusters and between-stratum variation in the intervention effect. Natural variation between clusters is not estimated separately but is still allowed for in the analysis. However, this approach will be restricted to studies with the relatively large number of cluster pairs required to estimate variance components with reasonable precision.[62]

Example of pair-matched cluster randomised design: the British Family Heart Study[57,63]

This study was designed to determine whether cardiovascular screening and lifestyle intervention in general practice achieved changes in coronary risk factors. Pairs of general practices were recruited in each of 13 towns, and one practice in each pair was randomly allocated to the intervention. Practices were matched for town, practice size and socio-demographic characteristics. The intervention consisted of a nurse-led programme using a family-centred approach with follow-up according to degree of risk. In men the reduction in risk score was 16% (95% confidence level 11–21%) at 1 year. The result was discussed in relation to the government's policy for screening in general practice.

Example of pair-matched non-equivalent group design: the Minnesota Heart Health Program[54]

This study evaluated an education program designed to promote change in coronary heart disease risk factors and behaviours across whole communities. Six communities were identified in Minnesota and North and South Dakota and matched into pairs based on size of community, type of community and distance from Minneapolis–St Paul. Randomisation was rejected because it was considered that 'randomisation of three units provides little assurance of equality of pooled communities'. Allocation to the education intervention was then decided on the basis of various factors including the degree of isolation of the media network, the number of local government structures to be dealt with, and whether the community was in Minnesota or the Dakotas.

Stratified designs

Under the stratified design, two or more clusters from each of several strata are allocated to each intervention group. The design is exemplified by the Child and Adolescent Trial for Cardiovascular Health study.[64] As for the paired design the clusters are grouped into strata according to one or more variables which are related to the outcome, but the stratified design differs qualitatively from the matched-pairs design because there is **replication** of clusters within each intervention–stratum combination. It is therefore possible to obtain an estimate of between-cluster variation from a stratified community-wide trial as the cluster effect can be separated from both the intervention effect and the stratum effect.

Using the completely randomised design as the benchmark in their simulations, Klar and Donner[51] found that compared with the matched-pairs design there is little loss in power for stratified designs when strata are not related to outcome. Another advantage of the stratified design over the matched design is that for studies in which there are a large number of clusters relative to the number of confounding factors it is easier to construct meaningful strata.[11,51]

In the light of these advantages, Klar and Donner[51] suggest that the stratified design is underutilised in comparison to the matched-pairs design and should be considered more often. The stratified design is limited to studies with eight clusters or more (four clusters in each of two strata).

Example of a stratified cluster randomised design: WHO trial for the evaluation of a new antenatal care programme[65]

This trial is being carried out to see whether a programme of antenatal care which only includes items of care of proven effectiveness has similar outcomes to current standard care. Fifty-three clinics were randomised after stratification by country (four different countries) and clinic size (small, medium and large). Stratification according to country was thought to provide some control over confounding by between-country differences. Clinic size was used as an additional stratifying factor because it was thought to be a marker for a range of baseline factors such as women's risk factors, socioeconomic status and geographical location.

Design issues for follow-up

The evaluation of health interventions often necessitates repeated assessments of health status over time. For example, a study might aim to evaluate the reduction in prevalence of unhealthy behaviour or the increase in the adoption of a healthy habit in the population. A single post-test comparison is insufficient for this purpose and measurements must be taken in each group at least once before and once after an intervention has been implemented.[10] Several authors[3,8,10] have suggested that repeated observation of clusters is a useful method for increasing the power of a study even if it is not necessary from a design viewpoint. Allowing for the change in the outcome of interest following intervention is a more sensitive method of evaluation than simply comparing postintervention values. As long as excessive costs are not incurred, this can increase the efficiency of a study. The Minnesota Cancer Pain Project[66] is an example of a community trial that used before and after intervention assessments.

Two designs can be used to sample individuals within clusters in cluster-based studies with follow-up. Under the **repeated cross-sectional** design a new sample of individuals is drawn from each of the clusters at each measurement occasion. Under the **cohort** design data are collected from the same sample of individuals at each measurement occasion. When the ratio of sample size to population size is high, a cohort will be generated within the framework of a repeated cross-sectional design as individuals may be sampled repeatedly by chance.[10] This overlap of repeated cross-sectional samples may be engineered deliberately, and some community intervention studies have set out to achieve this mixed design.[10,67] Mixed longitudinal designs can be costly and are not always feasible,[68] so a choice usually has to be made between repeated cross-sectional and cohort designs.

Choice of design: cohort or repeated cross-sectional sampling

If the main objective is to determine how an intervention changes individual level outcomes then a cohort design should be used. Cohort studies provide the opportunity to link individual outcomes directly to individual level prognostic factors.

For example, the Da Qing study of diabetes and impaired glucose tolerance[48] (see earlier) used a cohort design to study the effect of intervention by means of diet and exercise on the risk of developing diabetes.

When the main objective is to determine how an intervention affects some community level index of health, then a repeated cross-sectional design should be used, as it will generate data which are representative of the study communities throughout the study period. For example, the evaluation of Heartbeat Wales[45] (see earlier) studied the effect of a health promotion campaign on the prevalence of coronary heart disease risk factors. Cohorts are subject to ageing, migration, death, loss to follow-up and other factors which affect representativeness.[6,10,15] In choosing the most suitable design for follow-up, the aim of the intervention needs to be clearly defined so that the appropriate choice is made. This design choice is important because it will influence the validity of the study.

When both cohort and repeated cross-sectional designs are considered equally appropriate and there are no biases, cohort designs will generally offer greater statistical power.[6,10,15,56,69] In a cohort study,

measurements made over time on the same subject are often correlated and the estimate of the intervention effect has a smaller variance than that calculated from repeated cross-sectional data. The greater the correlation between measurements made on the same individual the smaller the variance of the intervention estimate and the more powerful the cohort design will be in relation to the repeated cross-sectional one. The study of the same individuals also controls for any unforeseen cohort effects.

In practice the cohort design will be more susceptible to bias than the repeated cross-sectional design,[10] and the power advantage diminishes as the length of follow-up increases because responses of the cohort members are time-dependent and consequently are less correlated as the study goes on. For the cohort design to have greater overall efficiency, the increase in precision needs to be sufficiently large to offset the potential bias. Cohort designs are generally best used in short trials where there is a high correlation between repeated measurements on the same subject. For studies in which the clusters are small in size, use of the cohort design may be advisable if overlapping cross-sectional surveys are to be avoided.[36] Repeated cross-sectional sampling is generally preferable in longer studies of larger clusters. Several researchers have introduced techniques which can be used at the planning stage of a study to establish which design will be more efficient.[67,69–71] Costs may also be taken into account.[72]

Sources of bias in studies with follow-up

- **Non-response**. Subjects who decline to participate in a trial may have different characteristics to those who do participate, rendering the sample less representative of the target population and introducing bias. Cohorts may be more vulnerable to this type of bias than repeated cross-sectional samples, because in order to follow up subjects it is more likely that they will have to reveal their identity and contact details.[10]
- **Attrition**, or loss to follow-up, only affects cohorts, and is particularly relevant when evaluation spans a long period. If a cohort member leaves a trial it is generally impossible to obtain further data. The characteristics of the drop-outs may be very different from those who are followed up, rendering the remaining subjects less representative of the original cohort. Whilst oversampling at baseline will guarantee a given level of precision it will not solve the problem of bias caused by attrition. Gillum and co-workers[73] devised a sample size formula to allow for loss to follow-up and Green and colleagues[15] presented methods for dealing with loss to follow-up at the analysis stage of a community-wide evaluation.
- **The testing or learning effect** is a specific type of **Hawthorne effect**. The problem is that the act of questioning may alter the behaviour of the subjects under study. Asking a group of respondents questions about the same health subject over a period of time could influence their behaviour by making them more sensitised to the topic.[5,10] Provided the population is large enough this is not a problem with repeated cross-sectional samples, but it can affect cohorts in a significant way. There is evidence to suggest that testing effects can also lead indirectly to differential non-response rates and bias between the intervention groups at follow-up surveys. Salonen and colleagues[36] suggested that higher non-participation rates at follow-up can be expected amongst those who have not yet been influenced by an intervention than amongst those who have. This may be of particular importance in non-blinded studies. If many patients who have not been influenced by an intervention drop out, the observed difference between groups may be exaggerated.
- **Maturation**, or ageing, affects only the cohort design.[6,10] In a study that runs over an appreciable period, the age distribution of the cohort will not be representative of the age distribution in the clusters for the later surveys. This can be detrimental to the external validity of the study if the outcome is age related.
- **Cross-contamination**. If mobility of the study subjects from cluster to cluster is high, then cross-contamination is possible and as a consequence the exposure status of the subjects may become unclear.[10] Ensuring that the intensity of the intervention is high enough to produce a clear and distinct measurable difference between the groups helps to diminish this problem.[36]

Notwithstanding the issues affecting the representativeness of the sample, internally valid conclusions about the effect of the intervention can still be reached if the groups are biased in the same manner.[67,74] Minimising the imbalance between groups with respect to confounders both at baseline and throughout the study and consistent implementation of the intervention will help to achieve internal validity.

Comparing the efficiency of cohort and repeated cross-sectional designs

Methods for comparing the efficiency of cohort and repeated cross- sectional designs fall into two categories.[67] The first is concerned with the power or precision of the estimate and the second with the

accuracy or overall efficiency of the estimate. Diehr and co-workers[67] pointed out that the two approaches appear to lead to the same general conclusion, but they can produce quite different results when the estimate of the intervention effect is biased. When bias is present it is more appropriate to assess the relative accuracy of the two designs because it takes into account both power and bias.

Diehr and co-workers[67] devised formulae for comparing the repeated cross-sectional design to the cohort design; the approach uses the **accuracy** criterion. The mean square error of the intervention estimate under the cohort design is compared with that obtained under repeated cross-sectional sampling using some 'gold standard' intervention estimate. Data from previous mixed design follow-up studies in which there are both cohort and repeated cross-sectional elements are required, which is one of the disadvantages of the approach. Such data will be hard to obtain.

Feldman and McKinlay,[71] using precision as their criterion, devised a formula for comparing the efficiency of cohort and repeated cross-sectional sampling for studies in which replicated measurements are made on the subjects once at baseline and once at follow-up. The method works on the premise that the cohort and repeated cross-sectional designs will be equally representative of the population of interest, and thus unlike the approach of Diehr and co-workers[67] it does not take account of bias. Their analysis of variance framework, from which the relative efficiency formula was derived, is also suited to **overlapping samples**, which have elements of both major designs.

Assume we have a study design for a long-term cluster-based intervention with replicated measurements (more than one measurement of the outcome for each subject at each measurement occasion) taken at baseline and follow-up for one intervention and one control group. The estimate of the intervention effect or the difference between changes in the intervention and control groups is given by

$$d = \sum_{it} \gamma_{it} y_{it\ldots} \qquad (1)$$

where $y_{it\ldots} = (1/JKM)\sum_{jkm} y_{itjkm}$ is the mean of the responses in the ith group at the tth time of measurement over all clusters, individuals and replicated measurements on individuals where there are J clusters each group, K individuals in each cluster and M measurements made per subject at each measurement occasion. Equal numbers of clusters within each group, equal cluster sizes and equal numbers of replicated measures per individual are assumed.

The γ_{it} are appropriate **contrast coefficients** which convert the term d into an expression that quantifies the change from baseline to follow-up in the intervention group relative to the control group. The contrast coefficients are governed by the following constraint:

$$\sum_i \gamma_{it} = \sum_t \gamma_{it} = 0 \qquad (2)$$

With two time points, the two contrast coefficients at each time point should be 1 and –1, allocated to the treatment and control groups, respectively, and likewise the two contrast coefficients for each group should be 1 and –1, allocated to the two time points follow-up and baseline, respectively. The contrast coefficients should be orthogonal.

Under the cohort design the variance of the estimator of the intervention effect, d, is given by

$$V_{\text{cohort}}(d) = (\sum_{it}\gamma_{it}^2)\left(\frac{(1-\rho_C)\sigma_C^2}{J} + \frac{(1-\rho_S)\sigma_S^2}{JK} + \frac{\sigma_E^2}{JKM}\right) \qquad (3)$$

ρ_C represents the autocorrelation between mean values for the same cluster over different time points, and ρ_S represents the autocorrelation between mean values for the same individual over different time points, σ_C^2 represents the random cluster variation, and is the sum of the time variant ($(1-\rho_C)\sigma_C^2$) and time-invariant ($\rho_C\sigma_C^2$) components of the cluster effect, σ_S^2 represents the random individual or subject level variation, and is the sum of the time variant ($(1-\rho_S)\sigma_S^2$) and time-invariant ($\rho_S\sigma_S^2$) components of the subject effect, and σ_E^2 represents the random variation between replicated measurements made on the same individual.

Under the repeated cross-sectional design the variance of the estimator, d, is given by

$$V_{\text{cross-sectional}}(d) = (\sum_{it}\gamma_{it}^2)\left(\frac{(1-\rho_C)\sigma_C^2}{J} + \frac{\sigma_S^2}{JK} + \frac{\sigma_E^2}{JKM}\right) \qquad (4)$$

$\rho_S = 0$ because a fresh sample of individuals is used at follow-up.

The formula for the efficiency of the repeated cross-sectional design relative to the cohort design is given by the ratio of these variances (3) and (4).

$$RE_{\text{cross-sectional}} = \frac{KM(1-\rho_C)\sigma_C^2 + M(1-\rho_S)\sigma_S^2 + \sigma_E^2}{KM(1-\rho_C)\sigma_C^2 + M\sigma_S^2 + \sigma_E^2} \quad (5)$$

When the expression is greater than unity the repeated cross-sectional design will yield a more precise estimate of intervention effect.

As part of this framework, Feldman and McKinlay[71] also provide formulae for calculating the relative size of the cohort design to the repeated cross-sectional that is required in order to achieve the same level of precision. The relative number of clusters per group is given by

$$J_{\text{cohort}} = \frac{KM(1-\rho_C)\sigma_C^2 + M(1-\rho_S)\sigma_S^2 + \sigma_E^2}{KM(1-\rho_C)\sigma_C^2 + M\sigma_S^2 + \sigma_E^2} J_{\text{cross-sectional}} \quad (6)$$

where J_{cohort} is the number of clusters per group in a cohort study that is required to achieve the same precision as $J_{\text{cross-sectional}}$ clusters per group in a repeated cross-sectional study, K is the fixed number of subjects per cluster and M is the fixed number of replicated measurements per individual.

The corresponding formula for the relative numbers of individuals per cluster is

$$K_{\text{cohort}} = \frac{M(1-\rho_S)\sigma_S^2 + \sigma_E^2}{M\sigma_S^2 + \sigma_E^2} K_{\text{cross-sectional}} \quad (7)$$

where K_{cohort} is the number of individuals per cluster in a cohort design that is required to achieve the same precision as $K_{\text{cross-sectional}}$ individuals per cluster at each time point in a repeated cross-sectional design, where M is the fixed number of replicated measurements per individual at each time point. The number of clusters per group is assumed to be fixed.

Feldman and McKinlay[71] also devised formulae for calculating the change in the number of clusters that is required to maintain the same precision given a change in the number of individuals per cluster, and vice versa. The formula for the number of clusters per group required to compensate exactly in terms of precision for a change in cluster size from K_1 to K_2 is

$$J_2 = \left(\frac{(1-\rho_C)\sigma_C^2 + (1-\rho_S)\sigma_S^2/K_2 + \sigma_E^2/K_2M_2}{(1-\rho_C)\sigma_C^2 + (1-\rho_S)\sigma_S^2/K_1 + \sigma_E^2/K_1M_1} \right) J_1 \quad (8)$$

given that J_1 clusters was previously sufficient.

The formula for the additional number of individuals per cluster required to compensate exactly in terms of precision for a change in the number of clusters per group from J_1 to J_2 is

$$K_2 = \left(\frac{J_1(1-\rho_S)\sigma_S^2 + J_1\sigma_E^2/M_2}{K_1(J_2-J_1)(1-\rho_C)\sigma_C^2 + J_2(1-\rho_S)\sigma_S^2 + J_2\sigma_E^2/M_1} \right) K_1 \quad (9)$$

given that a cluster size of K_1 was previously sufficient.

Formulae (8) and (9) are applicable to both cohort and repeated cross-sectional studies. Further, the formulae for the variance of the estimate (3) and (4) can be adapted to calculate sample sizes for given levels of precision. So Feldman and McKinlay's model is a useful framework for planning evaluations with follow-up, given prior knowledge from previous studies that are similar to the one being planned. The approach was formulated for studies with replicated measurements on individuals, but formulae could just as well be derived for studies with one measurement per individual at each time point.

An important requirement for comparing the efficiency of the cohort and repeated cross-sectional designs is that the cost of selecting and interviewing the subjects for the two designs be taken into account. McKinlay[72] modified Feldman and McKinlay's approach to take cost into account.

Examples of observational designs

The cohort and cross-sectional designs which were described above may also be used in observational healthcare evaluations. Black[75] outlined several reasons why observational studies should be used in evaluating the effectiveness of health care. The problems of making causal inferences from observational data were discussed by Bradford Hill, and the use of causal criteria was reviewed recently.[76]

Example of a cross-sectional observational design: hospital admissions for asthma in East London – associations with characteristics of local general practices, prescribing and population[77]

A survey was carried out of all 163 general practices in East London and City Health Authority. Hospital admission rates for asthma in each practice were related to selected practice characteristics. Higher asthma admission rates were found to be associated with smaller partnership size and with higher rates

of night visiting. The authors concluded that smaller practices should be helped to improve asthma care.

Example of a cohort observational design: survival with bladder cancer and characteristics of hospital or surgeons[78]

A cohort of patients registered with bladder cancer in the South Thames region was followed up for mortality. Survival analyses were carried out to determine whether the survival of patients treated at teaching hospitals differed from those treated at district hospitals and whether patients treated by urologists survived longer than those treated by general surgeons.

Ethical considerations

Ethical guidelines for the conduct of randomised trials are well established.[79] In general, each individual subject in a study should give informed consent to participate. In organisation- or area-based evaluations it may be difficult to obtain individual consent if the unit of allocation is large. For example, it would not usually be possible to obtain consent from all of the patients registered with a general practice, or from the residents of a health authority. In this situation, consent should be obtained at area or organisation level, perhaps from the GP or from senior officers of the health authority.[65,79] Such consent should usually be informed by wide consultation.[79] Edwards and colleagues recommend that it would not usually be ethical to remove individual choices that had previously existed without individual consent.[79] Donner[65] provided an illustration of this principle with reference to a trial of antenatal care. Clinics were randomised to standard care or to a new antenatal care schedule which only included items of care which were of proven value. Consent was given at clinic level, but women in the intervention arm were also asked to give informed consent. It was not considered necessary to obtain informed consent from women receiving standard antenatal care. Donner points out that this approach is similar to the randomised consent design proposed by Zelen.[65]

Summary

The study designs discussed in this chapter are summarised in *Table 3*.

TABLE 3 *Summary of design recommendations*

Design characteristics	Recommendation
Observational designs	Suitable for evaluating existing services
Intervention group with after only observations	Not often interpretable
Intervention group with before-and after observations	Not often interpretable
Intervention group and non-equivalent control group, after-only observations	Not often interpretable
Intervention and non-equivalent control group, before and after observations	Preferred design if randomisation not feasible. Restricted allocation recommended to reduce bias. Analysis controlling for confounders advisable
Interrupted time series design	May give interpretable results with a single group, inclusion of control group recommended
Randomised designs	Recommended
Restricted allocation	Recommended when the number of clusters is small, and when stratifying variables are associated with the outcome. Limitations of the matched pairs design are not widely appreciated
Cohort designs	More appropriate for short term studies of individual outcomes. Bias likely to increase with duration of study
Cross-sectional designs	More appropriate for studies of community level health indicators. May be less susceptible to bias but with lower power

Chapter 4

Measures of between-cluster variation

Introduction

In studies at the individual level, variation is between individuals. In cluster level studies, because of the correlation of individual level responses within clusters, there are two separate components of variation, within-cluster variation and between-cluster variation. Variation in the outcome among clusters will be larger than the variation among individuals within clusters.

Standard statistical methods for sample size estimation and analysis do not recognise the between-cluster component of variation in the outcome, and consequently cannot be applied directly at the individual level in cluster-based studies. Standard sample size calculations will lead to sample sizes that are too small. Standard methods of analysis will usually lead to confidence intervals that are too narrow and *p* values that are too small.

Appropriate methods for sample size calculation and analysis need to allow for within- and between-cluster variation. This section outlines methods that may be used to estimate within- and between-cluster components of variance, and subsequently describes how these are used to estimate the intraclass correlation coefficient and the design effect, which are used to adjust standard sample size calculations for cluster-based studies.

Estimating within- and between-cluster components of variance

Within-cluster and between-cluster variance components may be estimated by decomposing the variation in subject responses into these two constituent components using analysis of variance. Variance components may be estimated from a clustered sample through implementing a one-way analysis of variance of the outcome with the clustering variable as a random factor. The algebraic form of the appropriate model is

$$y_{ij} = \mu + \beta_j + e_{ij} \qquad (10)$$

where y_{ij} is the response of the ith individual within the jth cluster, μ is the mean of the responses, β_j is the random effect of the jth cluster and e_{ij} is the random error component. The β_j are independently and identically distributed (Gaussian) with zero mean and constant variance, σ_b^2. The e_{ij} are independently and identically distributed (Gaussian) with zero mean and constant variance σ_w^2.

Estimates of the between- and within-cluster variance components are extracted from the resulting analysis of variance table:

$$\hat{\sigma}_b^2 = (\mathrm{MSB} - \mathrm{MSW})/n_0 \qquad (11)$$

$$\hat{\sigma}_w^2 = \mathrm{MSW} \qquad (12)$$

where MSB is the between-cluster mean square and MSW is the within-cluster mean square, n_0 is the average cluster size calculated using

$$n_0 = \frac{1}{J-1}\left(N - \frac{\sum n_j^2}{N}\right) \qquad (13)$$

where J is the number of clusters, N is the total number of individuals, and n_j is the number of individuals in the jth cluster. This version of the average cluster size, n_0, is used because use of the arithmetic mean cluster size can lead to underestimation of the between-cluster variance component when the cluster sizes vary. n_0 approaches $\bar{n}$ when there is a large number of clusters.[80]

The analysis of variance approach can be extended to more complex designs, which include the other categorical and continuous variables that are important in the design of the study from which variance components have to be estimated. Thus there may be more than one intervention group, several strata, and several time points at which measurements are made. Interaction terms may also be needed, particularly if the data are from a study in which repeated measurements are made over time.[69,81] The clustering variable will generally be the only random factor in the model, but sometimes there may be a need to control for more than one level of clustering, in which case a more complex model with several random effects may be required. All interaction terms that involve a random main effect are themselves treated as random, so for instance the term for the cluster by time interaction effect in a

study with follow-up is random. Further details are provided in standard texts.[82]

The intraclass correlation coefficient (ρ)

The within-cluster and between-cluster components of variance may be combined in a single statistical measure of between-cluster heterogeneity (or within-cluster homogeneity), the intraclass correlation coefficient, ρ.

A more technical definition of ρ is the proportion of the true total variation in the outcome that can be attributed to differences between the clusters:

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2} \qquad (14)$$

where σ_b^2 is the between-cluster variance component and σ_w^2 is the within-cluster variance component. Equations (11) and (12) can be substituted into equation (14) and rearranged to show that ρ[83] is given by

$$\hat{\rho} = \frac{\text{MSB} - \text{MSW}}{\text{MSB} + (n_0 - 1)\text{MSW}} \qquad (15)$$

If individuals within the same cluster are no more likely to have similar outcomes than individuals in different clusters then the intraclass correlation will be 0. Conversely, if all individuals in the same cluster are identical with respect to the outcome, then the intraclass correlation is 1. In the context of cluster-based evaluations, ρ will usually assume small positive values. Negative values can be attributed to sampling error.

The design effect

In order to take account of between-cluster variation when estimating the sample size or carrying out the analysis, the variance term in standard statistical formulae for hypothesis testing or sample size calculation needs to be increased by the **design effect**. Kish[84] and Moser and Kalton,[85] in their works on survey sampling, define the design effect (Deff) as the ratio of the variance of the estimated outcome under the cluster sampling strategy (σ_c^2) to the variance that would be expected for a study with the same number of individuals using simple random sampling (σ_{srs}^2):

$$\text{Deff} = \frac{\sigma_c^2}{\sigma_{srs}^2} \qquad (16)$$

The design effect has also been referred to as the variance inflation factor.[12] In evaluations of cluster-based interventions the design effect will usually be greater than unity due to the presence of between-cluster variation. An interpretation of the design effect is the number of times more subjects a cluster-based evaluation should have compared with one in which individuals are randomised, in order to attain the same power. Estimates of sample size obtained using standard formulae should be multiplied by the design effect to ensure that enough subjects are included in the study. Further, individual level analyses need to allow for the design effect so that the effective sample size on which inferences are based is recognised.

It can be shown that the design effect is given by

$$\text{Deff} = 1 + (n - 1)\rho \qquad (17)$$

where n is the average cluster size and ρ is the intraclass correlation coefficient of the outcome.[84,85] This represents a more convenient and frequently used formula. It can be seen that the standard sample size calculation should be multiplied by a factor lying between unity and the average cluster size. The effective sample size for individual level analysis lies somewhere between the total number of individuals and the total number of clusters.

Given that the intraclass correlation coefficient of outcome is positive, equation (17) shows that the larger the average cluster size, the larger will be the design effect. In practice, ρ tends to be larger for smaller clusters,[86] but for large clusters, such as health authority populations, design effects will often be large even with small intraclass correlations. ρ is more generalisable than the design effect because it is independent of the number of individuals that are sampled from within each cluster and can therefore be readily compared across studies of similar design and purpose. Estimates of ρ can then be used to calculate the design effect of proposed studies at the planning stage. The following subsections provide further details of methods for obtaining estimates of ρ and variance components using analysis of variance for binary responses, studies with more than one intervention group, stratified cluster-based designs and studies with more than one level of clustering. The last two subsections respectively look at methods used to calculate confidence intervals for ρ and computer packages that can be used to estimate the ρ.

Estimating ρ from binary data

The analysis of variance model can also be used to calculate ρ when the outcome is dichotomous. However, McCulloch[87] warned that when covariates are controlled for, the analysis of variance framework is not applicable. Katz and colleagues[88] go further to claim that ρ itself is a poor measure of within-cluster association for binary responses. We conclude that the analysis of variance approach may be used, but methods for constructing confidence limits about ρ are not appropriate because distributional assumptions of the method are unlikely to be valid. The within-cluster dependence of binary responses tends to be low if the prevalence of the outcome is low, and outcome measures with a prevalence of 50% will lead to larger design effects.[89]

Analysis of variance methods generally require individual level data. When binary responses are aggregated at cluster level it is usually more convenient to estimate the kappa coefficient. Kappa is itself an intraclass correlation coefficient.[90] Fleiss[91] gives details of methods for estimating kappa. When the clusters are equal in size, Fleiss's kappa is algebraically equivalent to ρ as calculated for a binary outcome.[91] When the clusters are unequal in size, kappa will produce a value which is almost identical value to ρ,[86] particularly as the number of clusters increases. This is the case even when the cluster sizes are quite variable. The arithmetic mean cluster size, $\bar{n}$, is commonly used to estimate kappa whereas ρ is usually calculated using n_0, a weighted version of the mean cluster size (see earlier).

Katz and Zeger[92] also recommend the calculation of pairwise odds ratios to calculate the design effect using the method of alternating logistic regression.[93] The appropriate formula for calculating the design effect is given in Katz and Zeger.[92] Further discussion of methods for estimating ρ is given by Muller and Buttner.[94]

Estimating ρ from more than one group

A two-way analysis of variance is required to estimate ρ from more than one group, and the appropriate model is

$$y_{ijk} = \mu + \beta_{jk} + \gamma_k + e_{ijk} \qquad (18)$$

where y_{ijk} is the response of the ith individual of the jth cluster nested within the kth group, μ is the mean of the responses, β_{jk} is the random effect of the jth cluster in the kth group, γ_k is the fixed effect of the kth group relative to the mean, μ, and e_{ijk} is the random error component. The β_{jk} are independent and identically distributed (Gaussian) with zero mean and constant variance σ_b^2, and the e_{ijk} are independent and identically distributed (Gaussian) with zero mean and constant variance σ_w^2.

The intraclass correlation coefficient estimate is given by

$$\hat{\rho} = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_w^2} = \frac{\text{MSB} - \text{MSW}}{\text{MSB} + (n_0 - 1)\text{MSW}} \qquad (19)$$

where all terms are as previously described except n_0, which is now

$$n_0 = \frac{N - \sum_{k=1}^{K}\left(\sum_{j=1}^{J_k} n_{jk}^2 / n_k\right)}{\sum_{k=1}^{K}(J_k - 1)} \qquad (20)$$

where N is the total number of individuals, K is the number of intervention groups, J_k is the number of clusters in the kth group, n_{jk} is the size of the jth cluster in the kth group and n_k is the total number of individuals in the kth group. This approach effectively pools the group-specific variance estimates and thus separates the fixed group effect from the random cluster effect.

This analysis of variance model specification makes the assumption that the intraclass correlation coefficient is the same in each group, or in other words that σ_w^2 and σ_b^2 are constant across groups. Bartlett's procedure[83] can be used to test σ_w^2 for homogeneity across groups, and a test based upon Fisher's transformation can be used to assess whether ρ is constant between groups for balanced data.

Estimating ρ from matched pairs and stratified designs

As discussed in the previous chapter, ρ cannot generally be estimated directly from matched pairs of clusters because between-cluster variation is confounded with both intervention and stratum effects.[51] This is not a problem for stratified designs in which there are at least two clusters within each intervention—stratum combination, enabling

natural variability between clusters to be separated from intervention effects and between-stratum variation. Estimates of ρ are obtained by averaging out the ρ values of each intervention–stratum combination. This can be achieved by adjusting for both group and stratum in the analysis of variance model. Usually the strata will be represented as fixed effects.

For the matched-pairs design there are three conditions under which the intraclass correlation can be calculated using analysis of variance: when the intervention is ineffective, when matching is ineffective and when the intervention effect is homogeneous across each matched pair.[51]

If the intervention is ineffective, then a pooled estimate across strata of the between-cluster variance can be made. This is obtained by modelling stratum as a fixed effect in the analysis of variance model. If there is in fact an intervention effect and this method is used, then the intraclass correlation will be overestimated.

If matching is ineffective, then ρ can be calculated for each intervention group and pooled to get an estimate. This is achieved by using the formula for calculating the intraclass correlation from more than one group (see earlier). The more effective the matching the more the intraclass correlation will be overestimated using this method.

For dichotomous outcomes, if the true difference in underlying rates is stable from pair to pair then the intraclass correlation coefficient ($\hat{\rho}_1$) is given by[51]

$$\hat{\rho}_1 = \frac{\text{MSI} - \text{MSW}}{\text{MSI} + (n-1)\text{MSW}} \qquad (21)$$

where the mean square between clusters (MSI) is given by

$$\text{MSI} = \frac{\sum_{j=1}^{J} n(d_j - \bar{d})^2}{2(J-1)} \qquad (22)$$

n is the cluster size, $\hat{p}_{jk}$ is the event rate for the jth cluster in the kth group, and J is the number of pairs. The difference between cluster-specific event rates for the jth pair, d_j, is given by $d_j = \hat{p}_{j1} - \hat{p}_{j2}$. The mean of the differences between event rates, $\bar{d}$, is

$$\bar{d} = \sum_{j=1}^{J} d_j / J$$

The mean square within clusters is given by

$$\text{MSW} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{2} n\hat{p}_{jk}(1-\hat{p}_{jk})}{2J(n-1)} \qquad (23)$$

In the presence of treatment–stratum interaction, ρ will be overestimated. At least around 20 pairs of clusters are required for ρ to be estimated efficiently under the above formulae.

Estimating ρ for hierarchical designs

When calculating ρ for the level of clustering at which the intervention was implemented it is also necessary to control for all intermediate levels of clustering. For instance, in an organisation-based study in which general practices are randomised, the analysis of variance model will include both general practice and GP as random effects, the GP effect being nested within the general practice effect. Failure to allow for lower levels of clustering can lead to anomalous results, for example negative estimates of ρ.

Confidence interval construction for ρ

There are two main considerations when constructing confidence limits for ρ. Firstly, the clusters will typically be unequal in size, or **unbalanced**, which invalidates the appropriateness of the relatively simple formulae for balanced designs. Secondly, although analysis of variance produces an efficient point estimate of ρ for dichotomous outcomes, the method is not appropriate for estimating the corresponding confidence limits.

Exact confidence limits for ρ have been given by Searle[95] for the one-way balanced analysis of variance. The limits are

$$\left\{\frac{F/F_U - 1}{n + F/F_U - 1}, \frac{F/F_L - 1}{n + F/F_L - 1}\right\} \qquad (24)$$

where $F = \text{MSB}/\text{MSW}$, $\Pr\{F_L \le F \le F_U\} = 1 - \alpha$, $\Pr\{F_{k-1,\,k(n-1)} < F_U\} = \alpha_1$, $\Pr\{F_{k-1,\,k(n-1)} > F_L\} = \alpha_2$, $\alpha_1 + \alpha_2 = \alpha$ and, usually, $\alpha_1 = \alpha_2$. k is the number of clusters and n is the cluster size.

An exact method for calculating confidence limits for unbalanced data does exist,[96] but it is computationally intensive so approximate

methods tend to be used instead. Smith's method[97] is not best suited to cluster-based evaluations because it assumes a Normal distribution; this is unlikely to be the case for a small number of clusters. Donner and Wells[98] suggest that Searle's[95] exact method for balanced data should be adapted for the unbalanced design by replacing n in the formulae of the lower and upper limits by n_0. The approach is limited to studies in which only the random cluster effect needs to be modelled.

Computation of ρ and variance components

Several standard statistical packages can be used for estimating ρ but they may yield slightly different estimates when cluster sizes vary because not all packages use the same definition of average cluster size. The statistical packages SAS® and Minitab® include random effects analysis of variance procedures which can be used to estimate variance components, and both use n_0 as the weighted mean cluster size. Analysis of variance procedures in Minitab permit the modelling of multiple random factors as well as fixed covariates and factors.[99] In the SAS package three procedures may be used to obtain variance components, **Proc Nested**, **Proc Varcomp** and **Proc GLM**.[100] Proc GLM, although computationally intensive, is the most versatile. Numerous random effects can be modelled while controlling for other categorical and continuous variables. Proc Nested is a fast procedure for estimating variance components, but only random effects may be included in the model. Proc Varcomp can be used to adjust for categorical but not continuous variables and is much slower than Proc Nested. The package Stata® includes two commands for one way analysis of variance, **loneway** and **l1way**; the latter is more appropriate for estimating ρ when the cluster size varies.[101]

In recent years a number of specialist random effects modelling packages have been developed specifically for regression analysis of hierarchical, or clustered data. These packages may also be used to estimate components of variance. The specialist random effects modelling packages are superior to conventional analysis of variance because the latter gives biased estimates when clusters are highly variable in size, or when the number of covariates is large (see chapter 6).

Chapter 5

Sample size and power

Introduction

When estimating the sample size it is important to take into account the planned method of analysis. If the intention is to analyse cluster level outcomes, then a conventional sample size calculation may be performed to estimate the required number of clusters, using an estimate of the variance of the cluster level outcome.

If the intention is to analyse individual level outcomes, estimation of the required sample size is less straightforward. Individual level outcomes tend to be correlated within clusters, and this has the consequence that a study in which clusters are allocated to intervention groups will be less powerful than one in which an equivalent number of individuals is randomised (see chapter 1). If individual level analyses are intended, it will usually be necessary to multiply the sample size estimated from standard formulae by the design effect (see chapter 4).

In practice, when designing cluster-based evaluations the investigator often needs to estimate both the required total number of individuals and the required number of clusters. A minimum number of clusters is required for certain analytical techniques and a minimum number of individuals per cluster is also required to yield sufficiently precise cluster-specific estimates. If the cluster sizes are predetermined, the required number of clusters can be estimated by dividing the total number of individuals required by the average cluster size.

When it is feasible to sample individuals within clusters, the power of the study may be increased either by increasing the number of clusters or the number of individuals within clusters. Increasing the number of clusters rather than the number of individuals within clusters has several advantages.[10,61] Firstly, the result of the study will usually appear to be more generalisable if the intervention has been implemented in a number of different clusters. Secondly, a larger number of clusters allows more precise estimation of the intraclass correlation coefficient and thus more flexible approaches to analysis.[61] Thirdly, there is a limit to the extent to which power may be increased solely by increasing the number of individuals within clusters. The relative cost of increasing the number of clusters in the study, rather than the number of individuals within clusters, will be an important consideration when deciding on the final structure of the sample.

Obtaining appropriate estimates of ρ

Statistical formulae for sample size calculation usually incorporate the design effect (formula (17)). In order to carry out sample size calculations it is necessary to estimate the design effect from previous studies which are as similar as possible in terms of design to the one being planned. That is, studies that used similar-sized clusters containing similar types of individuals and used the same outcome. Having identified the appropriate studies, the intraclass correlation coefficients or between- and within-cluster variance components of the outcome must be estimated using methods outlined in chapter 4. A wide range of components of variance and intraclass correlations are provided in chapter 9 of this report.

Precise estimates of the intraclass correlation coefficient should be used, or at least the level of imprecision should be considered. ρ is estimated with some margin of error, and its precision is reduced with limited numbers of clusters and variable cluster sizes. It is sensible to take this imprecision into account and evaluate the effect that imprecision has on the sample size.[89] Feng and Grizzle[102] suggested that the results of studies of the size that yielded ρ should be simulated and a distribution of required sample sizes can then be generated. Hannan and colleagues[103] recommended that as intraclass correlations for large clusters are usually between 0 and 0.05, values in this range should be used to perform sensitivity analyses. Neither of these approaches actually provides confidence limits for the sample size, so where possible it may be advisable to use the confidence limits of the intraclass correlation coefficient to obtain corresponding confidence limits for the sample size.

Where appropriate, covariates other than design variables should be allowed for when estimating the intraclass correlation coefficient.[89,104–107] By controlling for covariates, particularly those at the cluster level, the sizes of ρ, the design effect

and subsequently the required sample can be reduced. In order to translate this into a power advantage the same covariates must also be controlled for when evaluating the intervention effect at analysis. For example, demographic characteristics such as gender, age and ethnic group are often controlled for. Murray and Short[105] observed that for alcohol-related outcomes, adjusting for cluster level covariates drastically reduced the size of the intraclass correlation coefficient. Raudenbush[107] also found that controlling for cluster level covariates has greater effect in reducing the intraclass correlation coefficient than controlling for covariates at the individual level. This might be expected as cluster level covariates will explain variation between clusters rather than within. Conversely, the addition of individual level covariates can increase the size of ρ if it reduces the size of the within-cluster variance component by a larger amount than it does the between-cluster component. Previous research is important in identifying the covariates that will improve the power of a study.

It may be worth calculating separate intraclass correlation coefficients for subgroups of important demographic variables, such as age and gender.[106] For example, if ρ is larger for women than men, then an estimate of ρ which is calculated on a sample of both sexes may underestimate the required sample size for women.

Estimating secular trends

Because follow-up intervention studies seek to alter the secular trend in the outcome of interest, sample size calculations should take into account the likely pattern of change in the absence of intervention. Spontaneous changes in the outcome are likely to occur as a result of demographic changes, cohort effects and non-specific temporal and period effects. Potential consequences of not accounting for trends in outcome or planning for spontaneous changes in trend are that the intervention will not be sufficiently intense to make a significant impact or the sample size may not be large enough to detect a relevant effect.[5,53,69] It is therefore necessary to estimate the secular trend in outcome as well as population means and standard deviations prior to the study.

Completely randomised design

Difference between group means

The **standard normal deviate** sample size formula for comparing two independent groups should be multiplied by the design effect.[108] The required number of individuals per group, n', is given by

$$n' = \frac{2(Z_\alpha + Z_\beta)^2 \sigma^2[1 + (n-1)\rho]}{d^2} = \frac{2(Z_\alpha + Z_\beta)^2 \sigma^2}{d^2} \text{Deff} \quad (25)$$

where Z_α is the value of the standardised score cutting off $\alpha/2$ of each tail, Z_β is the value of the standardised score defining a power level of $1 - \beta$, σ^2 is the variance of the outcome measure under simple random sampling, n is the expected cluster size or number of individuals sampled from within each cluster, ρ is the intraclass correlation coefficient, and d is the detectable difference between group means. This sample size formula is appropriate for studies with a large number of clusters. When the t test is required Z_α and Z_β can be replaced by $t_{2n'-2,\alpha}$ and $t_{2n'-2,\beta}$ respectively, with n' estimated iteratively.

When unequal cluster sizes are anticipated, Donner and colleagues[108] suggest the use of either the expected average cluster size, $\overline{n}$, or the more conservative expected maximum cluster size, n_{MAX} instead of n. The use of $\overline{n}$ can lead to underestimates of the required sample size for highly unbalanced designs if the size of each intervention group is less than around 100 individuals.[108]

Hsieh[19] derived sample size formulae for estimating differences between mean changes from baseline in studies with two repeated cross-sectional surveys, one before and one after intervention. The method is easily adapted for studies comparing means with one postintervention survey. Hsieh's formulae incorporate estimates of the variance components for the outcome (see chapter 4). The total required number of clusters, N, is given by

$$N = \frac{8(S_b^2 + S_w^2/m)(Z_\alpha + Z_\beta)^2}{d^2} \quad (26)$$

where S_b^2 is the estimate of the between-cluster component of variance of the outcome, S_w^2 is the estimate of the within-cluster component of variance of the outcome, m is the expected number of individuals per cluster, Z_α is the standard two-sided normal curve value with probability α, Z_β is the standard normal curve value with probability β and d is the detectable difference between the two intervention groups with respect to mean change from baseline.

Because the number of clusters will normally not be large enough to use the standard normal curve values, Hsieh suggests that the size of each group should be increased by one cluster for significance tests at the 5% level. Alternatively, t values can be used with N estimated iteratively, using

$$N = \frac{8(S_b^2 + S_w^2/m)(t_{\alpha;N-2} + t_{\beta;N-2})^2}{d^2} \quad (27)$$

By rearranging the above formulae the required number of individuals per cluster for a given number of clusters is

$$M = \frac{S_w^2}{Nd^2/8(t_{\alpha;N-2} + t_{\beta;N-2})^2 - S_b^2} \quad (28)$$

If these formulae are used for follow-up studies which use cohort sampling, then they will overestimate the required sample size because they do not adjust for the increased power of the study design resulting from the correlation between measurements made on the same individuals.

This method is adapted for studies with one cross-sectional measurement by halving the variation. The total number of clusters and the number of individuals in each cluster are respectively given by

$$N = \frac{4(S_b^2 + S_w^2/m)(t_{\alpha;N-2} + t_{\beta;N-2})^2}{d^2} \quad (29)$$

and

$$m = \frac{S_w^2}{Nd^2/4(t_{\alpha;N-2} + t_{\beta;N-2})^2 - S_b^2} \quad (30)$$

where d is now the detectable difference between group means.

Difference between group proportions

Under Cornfield's[109] approach, the sample size required for a study in which individuals are randomised is multiplied by an inflation factor equal to the ratio of the variance when clusters are randomised to the variance when individuals are randomised. The generalised formula for the inflation factor (IF) formula is[86,109]

$$\mathrm{IF} = \frac{\sigma_{\mathrm{prop}}^2 \bar{n}}{\bar{p}(1-\bar{p})} \quad (31)$$

where p_j is the proportion with the characteristic of interest in the jth cluster, J is the total number of clusters and n_j is the expected number of individuals sampled in the jth cluster. The proportion with the characteristic in the entire sample is

$$\bar{p} = \frac{\sum_{j=1}^{J} n_j p_j}{\sum_{j=1}^{J} n_j} \quad (32)$$

The variance of the cluster-specific proportions is

$$\sigma_{\mathrm{prop}}^2 = \frac{\sum_{j=1}^{J} n_j (p_j - \bar{p})^2}{\bar{n}^2 J} \quad (33)$$

and the mean cluster size is

$$\bar{n} = \frac{\sum_{j=1}^{J} n_j}{J} \quad (34)$$

Cornfield recommends that the inflated sample size should in turn be multiplied by a further correction factor (CF) in recognition that the true population variance of the cluster-specific proportions, σ^2_{prop}, is estimated with $J-1$ degrees of freedom. CF is calculated by dividing the t distribution score by the z value corresponding to significance level, α,

$$CF = \frac{t_{\alpha, J-1}}{z_\alpha} \quad (35)$$

where t_α is the t value defining the α level of significance with $J-1$ degrees of freedom, and z_α is the z value defining the α level of significance.

It is not necessary to quantify the within-cluster dependence for Cornfield's method, rather the variation in cluster-specific proportions should be known or estimated. Cornfield's method can be advantageous when dealing with large cluster sizes for which the prevalence of the phenomenon is available from routine statistical information. It is most appropriate when large units such as district health authorities are randomised because the estimates of the cluster proportions are likely to be reliable and obtainable. For smaller units such as general practices the data will often be unavailable and the method will be less effective.

Cornfield's inflation factor is equivalent to the design effect defined in chapter 4:

$$\frac{\sigma_{\mathrm{prop}}^2 n}{\bar{p}(1-\bar{p})} = 1 + (n-1)\rho \quad (36)$$

An alternative sample size method is given by the adaption by Donner and colleagues[29,108] of the standard sample size formula for comparing two proportions.[110] The required number of individuals per group, n', is

$$n' = \frac{(Z_\alpha + Z_\beta)^2[P_T(1-P_T) + P_C(1-P_C)][1 + (n-1)\rho]}{(P_T - P_C)^2} \quad (37)$$

$$= \frac{(Z_\alpha + Z_\beta)^2[P_T(1-P_T) + P_C(1-P_C)]}{(P_T - P_C)^2} \text{ Deff}$$

where Z_α is the value of the standardised score cutting off $\alpha/2\%$ of each tail, Z_β is the value of the standardised score cutting off the lower $\beta\%$ and defining a power level of $(100 - \beta)\%$, ρ is the intraclass correlation coefficient with respect to the binary trait, P_T is the expected event rate in the intervention group, P_C is the expected event rate in the control group and n is the cluster size. Kappa may be used as a measure of intraclass correlation as described in chapter 4.[91]

Matched-pairs design

When estimating the required sample size for matched-pairs designs it may be difficult to obtain appropriate estimates of between-cluster variation. This problem has been discussed recently by Donner and Klar[51] and Thompson and colleagues[61] (see chapter 3).

Difference between group means

Hsieh[19] derived sample size formulae for evaluating the differences between groups with respect to mean changes from baseline in studies with two repeated cross-sectional surveys. The required number of cluster pairs n_s is given by

$$n_s = 4(S_b^2 + S_w^2/m)(t_{\alpha;\, n-1} + t_{\beta;\, n-1})^2/d^2 \quad (38)$$

where S_b^2 is the estimate of the between-cluster component of variance of the outcome, S_w^2 is the estimate of the within-cluster component of variance of the outcome, m is the expected number of individuals per cluster, $t_{\alpha;\, n-1}$ is the t distribution value with two-sided probability α and $n - 1$ degrees of freedom, $t_{\beta;\, n-1}$ is the t distribution value with probability β and $n - 1$ degrees of freedom, and d is the detectable difference between mean changes from baseline. The solution can be found iteratively.

Rearranging the above formula the number of individuals per cluster is given by

$$m_s = S_w^2/[nd^2/4(t_{\alpha;\, n-1} + t_{\beta;\, n-1})^2 - S_b^2] \quad (39)$$

When there is only one cross-sectional survey, at follow-up, the respective formulae are

$$n_s = 2(S_b^2 + S_w^2/m)(t_{\alpha;\, n-1} + t_{\beta;\, n-1})^2/d^2 \quad (40)$$

and

$$m_s = S_w^2/[nd^2/2(t_{\alpha;\, n-1} + t_{\beta;\, n-1})^2 - S_b^2] \quad (41)$$

where d is now the detectable difference between group means.

Difference between group proportions

Hsieh's[19] formulae for comparing group means in pair matched designs may be used here with the differences between means replaced by differences between event rates. Shipley and co-workers[60] provide a formula for calculating the power of the cluster level paired t test for the matched-pairs cluster randomised design. The formula can be rearranged to calculate sample sizes for desired power and significance level. The formula, with the power given by the term $1 - \beta$, is

$$Z_\beta = \frac{|d|J}{[(2\bar{p} + d)/n_H + 2\sigma_{prop}^2(1-\rho)]^{1/2}(J+2)^{1/2}} - Z_\alpha \quad (42)$$

where Z_α is the two-sided normal curve value defining the level of significance of the test, Z_β is the normal curve value defining the power of the test, d is the detectable difference between proportions, J is the number of pairs of clusters in the study, $\bar{p}$ is the average value of the underlying event rates in the absence of any intervention, σ_{prop}^2 is the variance of the underlying event rates in the absence of any intervention, r is the correlation of the true underlying cluster-specific rates within each pair and n_H is the harmonic mean of the cluster sizes.

Stratified designs

Difference between group means

If techniques devised for completely randomised designs are employed here they are likely to overestimate the required sample size. An adapted version of Hsieh's[19] formula for the matched-pairs designs is recommended here.

Difference between group proportions, ratio of odds

Donner[111] presented formulae for calculating both the total number of subjects and clusters required to be randomised within strata to each intervention

group where cluster size is the stratifying factor. The total number of individuals required per group, N, is given by

$$N = (Z_\alpha T' + Z_\beta U')^2 / V^2 \quad (43)$$

where

$$T' = \tfrac{1}{2}\sqrt{\left(\sum_{j=1}^{J} t_j[1 + (m_j - 1)\rho][\bar{P}_j(1 - \bar{P}_j)]\right)} \quad (44)$$

(45)

$$U' = \sqrt{\left(\tfrac{1}{8}\sum_{j=1}^{J} t_j[1 + (m_j - 1)\rho][P_{j\mathrm{T}}(1 - P_{j\mathrm{T}}) + (P_{j\mathrm{C}}(1 - P_{j\mathrm{C}})]\right)}$$

$$V = \tfrac{1}{4}\sum_{j=1}^{J} t_j(P_{j\mathrm{T}} - P_{j\mathrm{C}}) \quad (46)$$

Z_α is the two-sided critical value of the normal distribution corresponding to the α% level of significance, Z_β is the critical value of the normal distribution corresponding to a power level of $(100 - \beta)$%, J is the number of strata, t_j is the fraction of individuals in the trial belonging to the jth stratum, m_j is the cluster size within the jth stratum, ρ is the intraclass correlation coefficient of the outcome, $\bar{P}_j$ is the proportion with the characteristic of interest in the jth stratum, and $P_{j\mathrm{T}}$ and $P_{j\mathrm{C}}$ are the success probabilities characterising the intervention and control group subjects, respectively, within the jth stratum.

The number of clusters to be assigned from the jth stratum to each intervention group is given by

$$n_j = (Nt_j)/(2m_j) \quad (47)$$

Donner's method allows the allocation of a constant number of subjects from each stratum to groups by setting all the stratum fractions, t_j, to be equal. Alternately, one may allocate a constant number of clusters from each stratum to the intervention groups:

$$t_j = m_j / \sum_{j=1}^{J} m_j \quad (48)$$

Advance knowledge is required of the stratum-specific success rates and ρ. This information may be available from previous studies.

Donner's method assumes that equal numbers of individuals and equal numbers of clusters are assigned from each stratum to each group. If cluster sizes are expected to vary within strata, then the formulae can be used with the mean anticipated cluster size in the jth stratum, $\bar{m}_j$, replacing m_j. The extent to which this will lead to an underestimate of the required sample size is dependent upon the variation in cluster size within stratum.

This method is a modified version of that used by Woolson and colleagues[112] for individual randomised studies. The technique is applicable to either Cochran's statistic (specifically) or the Mantel–Haenszel statistic for testing for a significant odds ratio; the two tests are essentially equivalent if the total number of individuals in each stratum is large.[111] This method may not be appropriate for the matched-pairs design because the method corresponds to an analysis which entails calculating the intraclass correlation coefficient from study data, which will not be possible without making special assumptions (see chapter 3).[111]

An example: sample size calculation allowing for clustering

To provide a simple numerical example we consider the planning of an audit of clinical care. Suppose an audit was to be carried out in the offices of single-handed GP and at clinics held in primary care health centres. We might propose to find out whether the proportion of all attenders who were taking antihypertensive treatment was the same in the two clinical settings.

Suppose it were assumed that about half (50%) of the health centre attenders and 40% of GP attenders were taking antihypertensive medication. How many subjects would be required to detect this difference? Using a standard sample size calculation with $\alpha = 0.05$ and a power of 0.80, if 408 subjects from each setting were included in the study then there would be sufficient power to detect a difference in case mix of this magnitude. However, because the subjects were to be sampled from a number of different clinics it would also be necessary to allow for between-cluster variation. We estimated that the average number sampled per clinic (in other words the cluster size) would be about 50. We did not know the intraclass correlation coefficient, ρ, so we carried out a sensitivity analysis using values of 0.01, 0.05 and 0.1. The results are shown in *Table 4*. After allowing for between-cluster variation, the sample size requirement was inflated to 608 if ρ was 0.01 or 2407 if ρ was 0.1. Thus, 12 clinics per group would suffice if ρ was 0.01, but 48 clinics per group would be needed if ρ was as high as 0.1. If the number sampled per clinic could be increased to 100, then eight clinics would be required if the ρ was 0.01, but 44 clinics would be needed if ρ was 0.1. In deciding on the final design of the study it would

TABLE 4 *Results of sample size calculation to detect a difference in proportions of 10%. Prevalence in control group = 50%, $\alpha = 0.05$, $1 - \beta = 0.8$*

Estimated ρ	Number sampled per cluster = 50			Number sampled per cluster = 100		
	Design effect per group	Individuals required per group	Clusters required	Design effect per group	Number required per group	Clusters required
0.00	1.00	408	8	1.00	408	4
0.01	1.49	608	12	1.99	812	8
0.05	3.45	1408	28	5.95	2428	24
0.10	5.9	2407	48	10.90	4447	44

be necessary to consider the feasibility and costs of increasing the number of clinics in the study rather than the number of individuals sampled per clinic.

This example shows how important it will be to have an approximate estimate of the intraclass correlation coefficient when designing a study.

Chapter 6
Analysis

Correlation between responses of individuals within clusters must be allowed for in the analysis of cluster level evaluations. Failure to do so will usually result in *p* values that are too small and confidence intervals that are too narrow. The choice of analytical approach may be restricted by the design of the study, and by the number of clusters and individuals included in the sample. Decisions made at the planning stage are therefore very important.

Alternative approaches to analysis

There are three basic approaches to analysis:

- cluster level analysis with the cluster means, proportions or log odds used as the observations
- univariate analysis of individual level data with standard errors adjusted for the design effect
- regression analysis of individual level data using methods for clustered data.

Cluster level analysis uses the cluster means, proportions or log odds as the observations, and applies standard parametric or non-parametric analytical methods. Individual level analysis uses the outcome values for individual subjects as observations. In order to allow for clustering of data, the design effect is incorporated into standard formulae. This approach involves estimating the intraclass correlation coefficient of the outcome from the study data, and about 20 clusters are required to calculate ρ with reasonable precision.[113] Adjusted individual level hypothesis tests often assume a common value for ρ across groups, and for this reason may be less applicable to non-randomised data.[113] Analysis of non-randomised designs will usually require controlling for baseline characteristics and this will necessitate use of multiple regression methods for clustered data. Further details of these different approaches are provided below, together with a worked example. We recommend that the general reader should look at the worked example before proceeding to the more detailed discussion of statistical methods included in the two sections that now follow.

Univariate cluster level and individual level tests

In this section, we outline univariate methods of analysis for each of the three main designs: completely randomised, pair matched and stratified designs. We present methods suitable for analysis of continuous and dichotomous outcomes. The methods described fall into two main groups: adjusted individual level and cluster level methods of analysis.

Adjusted individual level methods adapt standard statistical methods by incorporating the design effect into formulae for the standard error. They require a sufficiently large number of clusters to obtain a stable estimate for ρ. Approaches that assume a common value for ρ across groups will be less applicable to non-randomised data.

Cluster level analyses are performed using the cluster means, proportions or log odds as the observations with application of standard parametric or non-parametric analytical methods. For example, a two-sample *t* test could be performed on the cluster means. When the number of clusters in a study is too small to estimate the intraclass correlation coefficient, cluster level analyses may be preferred to analysis at the individual level. The *t* test is known to be robust when applied to studies with as few as three clusters per group. Individual level confounders cannot be included directly, but cluster level statistics may be standardised for variables such as age and sex. Cluster level analysis has the appeal that the unit of intervention and analysis are the same, but relationships demonstrated at the cluster level do not always hold at the individual level. This is sometimes referred to as the ecological fallacy.

Completely randomised design

Continuous outcomes

Donner and Klar[59] presented a formula for comparing two means in a cluster randomised study generalisable to studies in which the clusters differ in size. The test statistic is an adjustment of the normal deviate test as applied to the individual, and should be evaluated against the table of the normal distribution:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{[\hat{V}(\bar{X}_1 - \bar{X}_2)]^{1/2}} \quad (49)$$

where $\bar{X}_1 - \bar{X}_2$ is the difference between the group means for the intervention effect, and

$$\hat{V}(\bar{X}_1 - \bar{X}_2) = S_p^2 \left(\frac{C_1}{N_1} + \frac{C_2}{N_2} \right) \quad (50)$$

is the variance of $\bar{X}_1 - \bar{X}_2$, where

$$S_p^2 = \frac{\sum_{k=1}^{2} \sum_{j=1}^{J} (n_{jk} - 1) s^2_{jk}}{\sum_{k=1}^{2} \sum_{j=1}^{J} (n_{jk} - 1)} \quad (51)$$

is the overall pooled variance, n_{jk} is the number of individuals in the jth cluster of the kth group, s^2_{jk} is the cluster-specific variance for the jth cluster of the kth group,

$$C_k = 1 + \left[\left(\sum_{j=1}^{J} \frac{n^2_{jk}}{N_k} \right) - 1 \right] \hat{\rho} \quad (52)$$

is the group-specific variance correction factor, $\hat{\rho}$ is the intraclass correlation coefficient estimated from the study data, and N_k is the number of individuals in the kth group. The corresponding confidence interval for the difference between two means is

$$\bar{X}_1 - \bar{X}_2 \pm z_\alpha [\hat{V}(\bar{X}_1 - \bar{X}_2)]^{1/2} \quad (53)$$

where z_α defines the level of confidence with approximate two-sided $(1 - \alpha)100\%$ confidence limits.

Random effects meta-analysis offers an alternative approach to univariate analysis.[61] In this application of the technique, results are pooled across clusters rather than across studies. The approach may be used for both continuous and dichotomous outcomes in studies with and without stratification. To implement meta-analysis for studies without stratification, the cluster-specific means are pooled across clusters for each group and the difference between the two group-specific statistics is tested for significance.

When the number of clusters per group is small (usually less than ten clusters per group) it may be better to use the cluster means themselves as the analytic units and apply a **two-sample *t* test**.[114] A weighted t test can be used when clusters are unequal in size.

Dichotomous outcomes

The confidence interval for the risk difference between groups is given by[59]

$$(\hat{P}_1 - \hat{P}_2) \pm z_\alpha \hat{\mathrm{SE}}(\hat{P}_1 - \hat{P}_2) \quad (54)$$

where z_α defines the level of confidence with approximate two-sided $(1 - \alpha)100\%$ confidence limits. The standard error of the difference is given by

$$\hat{\mathrm{SE}}(\hat{P}_1 - \hat{P}_2) = \left(\frac{C_1 \hat{P}_1 \hat{Q}_1}{N_1} + \frac{C_2 \hat{P}_2 \hat{Q}_2}{N_2} \right)^{1/2} \quad (55)$$

where C_k is the variance correction factor for the kth group, and is given by equation (52), $Q_k = 1 - P_k$ and N_k is the number of individuals in the kth group. A test of significance can be performed by pooling the two samples to give the overall proportion.[110] The design effect is then incorporated into the formula for the **standardised normal deviate test**

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\mathrm{SE}(\hat{P}_1 - \hat{P}_2)} \quad (56)$$

An appropriate formula is shown as equation (9) of Donner and colleagues.[108]

The group-specific **adjusted chi-squared** approach,[113] an adjusted individual level method, is applicable to multigroup studies,[115] though the version of the formula below is for two groups. The technique is closely related to the adjusted method for constructing confidence intervals for the risk difference between two groups. The test statistic, χ^2_A, should be evaluated using the chi-squared distribution with one degree of freedom:

$$\chi^2_A = \sum_{i=1}^{2} \frac{n_k (\hat{P}_k - \hat{P})^2}{C_k \hat{P}(1 - \hat{P})} \quad (57)$$

where n_{jk} is the number of people in the jth cluster of the kth group, n_k is the number of individuals in the kth group,

$$\bar{n}_{Ak} = \sum_{j=1}^{J_k} n^2_{jk} / n_k$$

is the weighted average cluster size in the kth group, J_k is the number of clusters in the kth group, $\hat{\rho}$ is the intraclass correlation coefficient, $C_k = 1 + (\bar{n}_{Ak} - 1)\hat{\rho}$ is the clustering correction factor for the kth group, $\hat{P}_k$ is the event rate in the kth group and $\hat{P}$ is the overall proportion with characteristic of interest.

Clustering correction factors are computed for each intervention group, and it is assumed that there is no 'significant' difference between the design effects of each group. For this reason the method is not well suited to non-randomised studies. Again, the group-specific adjusted chi-squared method is not reliable when less than ten clusters are randomised to each group because the intraclass correlation coefficient cannot be estimated with sufficient precision.[113]

The **ratio estimate chi-squared** method is another adjustment of the chi-squared statistic but the concept of the design effect is treated in a different manner; it is calculated for each group separately by dividing the variance of the event rate in each group, under cluster randomisation by the variance under individual randomisation. The test is generalisable to studies with more than two intervention groups.

The following statistic should be referred to the chi-squared distribution with one degree of freedom:

$$\chi^2_{\mathrm{R}} = \sum_{k=1}^{2} \frac{\tilde{n}_k (\hat{P}_k - \tilde{P})^2}{\tilde{P}(1-\tilde{P})} \qquad (58)$$

where $\tilde{n}_k = n_k / d_k$, n_k is the number of individuals in the kth group, $\hat{P}_k$ is the event rate in the kth group, d_k is the design effect for the kth group given by

$$d_k = \hat{\mathrm{Var}}_{\mathrm{R}}(\hat{P}_k) / \hat{\mathrm{Var}}_{\mathrm{B}}(\hat{P}_k) \qquad (59)$$

$$\hat{\mathrm{Var}}_{\mathrm{R}}(\hat{P}_k) = J_k (J_k - 1)^{-1} n_k^{-2} \sum_{j=1}^{J_k} (X_{jk} - n_{jk}\hat{P}_k)^2 \qquad (60)$$

n_{jk} is the number of individuals in the jth cluster of the kth group,

$$\hat{\mathrm{Var}}_{\mathrm{B}}(\hat{P}_k) = [\hat{P}_k (1 - \hat{P}_k)] / n_k \qquad (61)$$

J_k is the number of clusters in the kth group, X_{jk} is the number of successes in the jth cluster of the kth group, $\tilde{X}_k = X_k / d_k$ is the effective number of successes in the kth group and

$$\tilde{P} = \sum_{k=1}^{2} \tilde{X}_k / \sum_{k=1}^{2} \tilde{n}_k$$

The ratio estimate chi-squared test is well suited to non-experimental comparisons as neither the design effect nor intraclass correlation coefficient are assumed to be the same in each group. However, the method is generally less powerful than the group-specific adjustment approach for experimental comparisons, and is really only applicable in very large samples because it is based on the use of between-cluster variation for the estimation of error. This approach will usually require **much more** than ten clusters per group to ensure validity, and will tend to work better when there is little variation in cluster size and the amount of clustering in the data is small.[113] A version of this test has been developed that does rely on a pooled estimate of the design effect.[116]

Donner and Klar[59] provide formulae for the confidence interval of the odds ratio estimate using what they refer to as the **clustered Woolf** method. Given that $\hat{\psi} = (\hat{P}_1 \hat{Q}_2)/(\hat{P}_2 \hat{Q}_1)$ is the odds ratio estimate, where $\hat{P}_k$ is the proportion with the characteristic in the kth group and $\hat{Q}_k = 1 - \hat{P}_k$, approximate two-sided $(1 - \alpha)100\%$ confidence limits are given by

$$\exp\left(\log_e \hat{\psi} \pm z_\alpha \left[\hat{V}(\log_e \hat{\psi})\right]^{1/2}\right) \qquad (62)$$

$$\hat{V}(\log_e \hat{\psi}) = \left(\frac{C_1}{N_1 \hat{P}_1 \hat{Q}_1} + \frac{C_2}{N_2 \hat{P}_2 \hat{Q}_2}\right) \qquad (63)$$

gives the variance of the log odds ratio estimate, C_k is the variance correction factor for the kth group given by equation (52) and N_k is the total number of individuals in the kth group. If $\hat{P}_k = 0$ or 1, then 0.5 needs to be added to both the numerator and denominator of $\hat{P}_k$ for the variance of the log odds ratio estimate to be defined. This method assumes a large number of clusters because it requires an estimate of ρ.

As for continuous outcomes, **random effects meta-analysis** is appropriate here but this time the cluster-specific log odds or proportions are pooled across clusters for each group, rather than the mean, and the difference between the two groups is then tested for significance.

As for continuous outcomes, Donner and Klar[114] and Klar and colleagues[80] recommended that when the number of clusters per group is less than around ten, a **two-sample *t* test** should be performed using the cluster proportions as observations. If inferences are to be made on the relative odds scale then the cluster-specific log odds should be used instead of proportions. A **weighted *t* test** is appropriate when the cluster sizes are variable. A disadvantage of using the two-sample *t* test is that it does not reduce to standard methods for testing the null hypothesis of equal event rates when there is no clustering effect.

The cluster level *t* test is reputed to be robust for proportions, and even when the assumptions are

violated the method can be used with as few as three clusters per group so long as variation in the cluster sizes is not large. In studies where the number of clusters is less than ten non-parametric tests can also be used,[113] though they suffer from the disadvantage of low statistical power. Among the non-parametric tests, **Wilcoxon's rank-sum test**, which uses the rank order of the cluster-specific event rates, and **Fisher's two-sample permutation test**, which takes account of the magnitudes of the event rates, may be used.[15,113] Both tests require at least four clusters per group in order to reach significance at the 5% level.[6]

Matched-pairs design

Klar and Donner have drawn attention to the limitations of the paired design.[51] The difficulty of separating the effect of the intervention from between-cluster variation within strata means that the analytical options are more restricted than for studies with strata containing more than two clusters (see chapter 3). However, Thompson and co-workers[61] showed that **random effects meta-analysis** is a valid approach for analysing data from matched-pairs cluster-based studies for which the between-cluster variation does not need to be estimated explicitly (see chapter 3). Under this approach the statistic of interest, stratum-specific differences in means for continuous outcomes, or differences in proportions or log odds for dichotomous outcomes, are calculated and pooled across strata.

Continuous outcomes

The **paired *t* test** applied at the cluster level using the cluster-specific means as observations is applicable here. Evidence suggests that the paired *t* test is robust enough to test for significance between group means even if there is substantial variation in cluster size from pair to pair.[58] However, the standard paired *t* test is not strictly valid if the cluster sizes within each pair are unequal, and the **weighted paired *t* test** should used be under such circumstances. It is also appropriate to use the weighted version of the paired *t* test when there is a small number of large clusters if the cluster size varies between strata. Thus, the weighted paired *t* test supersedes the standard paired *t* test in the context of this review.[58] The appropriate test statistic for the weighted *t* test is

$$T_{\mathrm{WP}} = \frac{\bar{d}_{\mathrm{w}} \sum_{j=1}^{J} W_j}{S_{\mathrm{dw}} \sqrt{\sum_{j=1}^{J} W_j^2}} \qquad (64)$$

which should be referred to tables of the *t* distribution with $J-1$ degrees of freedom, where J is the number of strata.

The corresponding confidence interval for the intervention effect is given by

$$\bar{d}_{\mathrm{w}} \pm \frac{\left(t_{\alpha/2} S_{d\mathrm{w}}^2 \sum_{j=1}^{J} W_j^2\right)^{1/2}}{\sum_{j=1}^{J} W_j} \qquad (65)$$

where

$$\bar{d}_{\mathrm{w}} = \sum_{j=1}^{J} W_j d_j / \sum_{j=1}^{J} W_j$$

is the weighted difference between the group means, the stratum weights are given by $W_j = (nj_1 nj_2)/nj_{\mathrm{T}}$, n_{jk} is the size of the jth cluster in the kth group, nj_{T} is the size of the jth stratum, d_j is the difference between the cluster-specific means in the jth stratum, $t_{\alpha/2}$ gives the two-sided $(1-\alpha)100\%$ confidence limits and

$$S_{d\mathrm{W}}^2 = \sum_{j=1}^{J} W_j (d_j - \bar{d}_{\mathrm{W}})^2 / \sum_{j=1}^{J} W_j$$

is the variance of $\bar{d}_{\mathrm{W}}$.

Random effects meta-analysis may be used here, pooling the stratum-specific statistic of interest across strata.

Dichotomous outcomes

Donner's[111] **adjustment to the Mantel–Haenszel test statistic** (see below) for stratified cluster-based studies, can only be used in the special case of the matched-pairs design if an estimate of ρ can be obtained. Non-parametric methods using cluster level observations such as Wilcoxon's signed rank test and Fisher's permutation test[58] have too little power for small numbers of clusters. This section concentrates on more suitable analytical approaches.

Random effects meta-analysis may be used here pooling the stratum-specific statistic of interest across strata.

The **paired *t* test** may be applied at the cluster level to the differences between proportions for pairs of clusters within strata.[58,59] Although the assumptions of equal variances of the stratum-specific risk differences and normality are not met,[58] the test is still fairly robust provided that there is a large number of matched pairs. The paired *t* test can also be used under these conditions when stratum size is variable but the weighted version described below is likely to be more efficient.

The **weighted paired *t* test** should be used when the cluster size within strata is not constant.[58,59]

It is well suited to studies with a small number of large clusters that vary in size. The appropriate test statistic should be referred to tables of the t distribution with $J-1$ degrees of freedom where J is the number of strata

$$T_{\mathrm{WP}} = \frac{\bar{d}_{\mathrm{W}}\Sigma W_j}{S_{\mathrm{dW}}\sqrt{\Sigma W_j^2}} \tag{66}$$

The other components of the equation are as defined for equation (64) for continuous outcomes except that the risk difference rather than the difference in means is used.

When analysing a small number of large matched clusters with fairly unbalanced designs Donner and Donald[58] recommend the use of the **weighted paired *t* test based on the empirical logistic transform of the crude event rates**. Along with the standard weighted paired t test this method is an alternative for studies in which cluster size within strata is not constant. Simulations have shown this method to be the best for comparing matched proportions in which a small number of large clusters is allocated to groups.[58,117] The appropriate test statistic is

$$t_{wl} = \frac{\bar{d}_{\mathrm{w}l}\Sigma w_{jl}}{S_{\mathrm{l}}\sqrt{\Sigma w_{jl}^2}} \tag{67}$$

to be referred to tables of the t distribution with $J-1$ degrees of freedom where J is the number of strata. The weighted mean log odds ratio over the strata is given by

$$\bar{d}_{wl} = \frac{\Sigma w_{jl} d_{jl}}{\Sigma w_{jl}} \tag{68}$$

the stratum-specific weights are given by

$$w_{jl} = \frac{1}{\mathrm{var}(d_{jl})} \tag{69}$$

the log odds ratio for the jth stratum is given by

$$d_{jl} = l_{j1} - l_{j2} \tag{70}$$

and the log odds for the jth cluster in the kth group is given by

$$l_{jk} = \log_{\mathrm{e}}\left(\frac{a_{jk}+\frac{1}{2}}{n_{jk}-a_{jk}+\frac{1}{2}}\right) \tag{71}$$

$$\mathrm{var}(d_{jl}) = \frac{(n_{j1}+1)(n_{j1}+2)[1+(n_{j1}-1)\hat{\rho}]}{n_{j1}(a_{j1}+1)(n_{j1}-a_{j1}+1)} + \frac{(n_{j2}+1)(n_{j2}+2)[1+(n_{j2}-1)\hat{\rho}]}{n_{j2}(a_{j2}+1)(n_{j2}-a_{j2}+1)} \tag{72}$$

$$S_l^2 = \frac{\Sigma w_{jl}(d_{jl}-\bar{d}_{wl})^2}{\Sigma w_{jl}} \tag{73}$$

where a_{jk} is the number of people with the characteristic of interest in the jth cluster of the kth group and n_{jk} is the total number of people in the jth cluster of the kth group.

The corresponding confidence limits for the intervention effect are given by

$$\bar{d}_{wl} \pm \frac{t_\alpha S_l \sqrt{\sum_{j=1}^{J} w_{jl}^2}}{\sum_{j=1}^{J} w_{jl}} \tag{74}$$

where t_α provides two-sided confidence limits of $(1-\alpha)100\%$.

Use of the **clustered Woolf odds ratio estimator** and associated confidence interval for the matched design is equivalent algebraically to the weighted paired t test based on the empirical logistic transform of the crude event rates. The only difference is the latter is used to make inferences on the risk difference rather than the relative odds scale. The clustered method is superior to the classical Woolf odds ratio estimator for making inferences in cluster-based studies provided there is a positive intraclass correlation coefficient.

The clustered Woolf odds ratio estimator is

$$\hat{\varphi}_{\mathrm{CW}} = \exp(\hat{\gamma}_{\mathrm{CW}}) \tag{75}$$

the weighted log odds ratio is given by

$$\hat{\gamma}_{\mathrm{CW}} = \frac{\sum_{j=1}^{J} W_{j\mathrm{C}}\hat{\gamma}_j}{\sum_{j=1}^{J} W_{j\mathrm{C}}} \tag{76}$$

the weights are given by

$$W_{j\mathrm{C}} = \left(\frac{C_{j1}}{n_{j1}\hat{P}_{j1}\hat{Q}_{j1}} + \frac{C_{j2}}{n_{j2}\hat{P}_{j2}\hat{Q}_{j2}}\right)^{-1} \tag{77}$$

the log odds ratio specific to the jth stratum is given by

$$\hat{\gamma}_j = \log_e(\hat{\varphi}_j) \tag{78}$$

the odds ratio for the jth stratum is

$$\hat{\varphi}_j = (\hat{P}_{j1}\hat{Q}_{j2})/(\hat{P}_{j2}\hat{Q}_{j1}) \tag{79}$$

and the clustering correction factor for the jth cluster in the kth group is given by

$$C_{jk} = 1 + (n_{jk} - 1)\hat{\rho}_P$$

where n_{jk} is the total number of subjects in the jth cluster of the kth group, J is the number of strata, $\hat{P}_{jk}$ is the estimate of the proportion with the characteristic of interest in the jth cluster of the kth group, $\hat{Q}_{jk} = 1 - \hat{P}_{jk}$, and $\hat{\rho}_P$ is the intraclass correlation coefficient estimated from the study data.[59]

The corresponding two-sided confidence limits about the odds ratio estimate, $\hat{\varphi}_{CW}$, are then given by

$$\exp\left[\hat{\gamma}_{CW} \pm t_{\alpha/2}(\hat{V}(\hat{\gamma}_{CW}))^{1/2}\right] \tag{80}$$

where $t_{\alpha/2}$ defines the two sided $(1-\alpha)100\%$ confidence limits and the variance of the weighted log odds ratio, $\hat{V}(\hat{\gamma}_{CW})$, is given by

$$\hat{V}(\hat{\gamma}_{CW}) = \frac{S_C^2 \sum_{j=1}^{J} W_{jC}^2}{\left(\sum_{j=1}^{J} W_{jC}\right)^2} \tag{81}$$

where

$$S_C^2 = \frac{\sum_{j=1}^{J} W_{jC}(\hat{\gamma}_j - \hat{\gamma}_{CW})^2}{\sum_{j=1}^{J} W_{jC}} \tag{82}$$

A test of significance corresponding to this confidence interval is obtained by referring $t = \gamma_{CW}/\sqrt{[\hat{V}(\gamma_{CW})]}$ to tables of the t distribution with $J-1$ degrees of freedom.[117] It is worth noting that the odds ratio estimate itself remains consistent in the presence of clustering.

Use of the clustered Woolf odds ratio estimate entails estimating ρ_P, and for this reason its use may be problematic for matched-pairs designs, particularly where there is a small number of clusters.

Stratified designs

Continuous outcomes

The **paired *t* test** can be used at the cluster level if the distribution of cluster sizes from stratum to stratum is similar; the observations are means calculated for each combination of stratum and group. If there is much variation between stratum sizes the **weighted paired *t* test** is preferable.[59] **Random effects meta-analysis** is another valid approach with the stratum-specific differences between means pooled across strata.[61]

Dichotomous outcomes

The analytical techniques that Donner and Klar presented for testing prevalence differences and odds ratios for significance under the matched-pairs design are appropriate for the stratified randomised design also;[59] the **paired *t* test** and **weighted paired *t* test** applied at the cluster level and the clustered Woolf method applied at the individual level. In contrast to the matched-pairs design the intraclass correlation can be estimated with ease but an estimate appropriate to the stratified design with multiple replication of clusters should be used.

Donner's[111] adjustment to the **Mantel–Haenszel chi-squared test** for a significant odds ratio estimate is also applicable here. The procedure is a generalisation of that given by Donald and Donner.[118] The following statistic should be referred to the chi-squared distribution with one degree of freedom.

$$\chi^2_{MHA} = \frac{\left(\left|\sum_{j=1}^{J} \frac{a_{jT}(n_{jC} - a_{jC}) - (a_{jC}(n_{jT} - a_{jT})}{n_{jT}B_{jC} + n_{jC}B_{jT}}\right| - \frac{1}{2}\right)^2}{\sum_{j=1}^{J} \frac{n_{jT}n_{jC}a_j(n_j - a_j)}{n_{jT}B_{jC} + n_{jC}B_{jT} - 1)\,n_j^2}} \tag{83}$$

where J is the number of strata, a_{jT} is the number of subjects with the characteristic of interest in the jth stratum of the intervention group, a_{jC} is the number of subjects with the characteristic in the jth stratum of the control group, $a_j = a_{jT} + a_{jC}$, n_{jT} is the total number of subjects in the jth stratum of the intervention group, n_{jC} is the total number of subjects in the jth stratum of the control group, and $n_j = n_{jT} + n_{jC}$.

The clustering correction factors, B_{jT} and B_{jC} for the intervention and control groups, respectively, are given by

$$B_{jT} = \frac{\sum_m m[1 + (m-1)\rho]N_{jTm}}{\sum_m mN_{jTm}} \tag{84}$$

and

$$B_{jC} = \frac{\sum_m m[1 + (m-1)\rho]N_{jCm}}{\sum_m mN_{jCm}} \quad (85)$$

where N_{jTm} is the number of treatment group clusters in the jth stratum having exactly m subjects, N_{jCm} is the number of control group clusters in the jth stratum having exactly m subjects, and ρ is the intraclass correlation coefficient.

In practice ρ will be unknown and is estimated from the data. It is obtained by implementing an analysis of variance as described in chapter 4. The adjusted Mantel–Haenszel procedure cannot generally be used for the matched-pairs design, where each stratum contains just two clusters, as the intraclass correlation cannot be estimated from the study. Thus, use of this method for matched-pairs studies will usually lead to conservative results.

Random effects meta-analysis is another valid approach with the stratum-specific differences between proportions or the odds ratio estimates pooled across strata.

Regression methods for clustered data

Limitations of standard regression methods

It is not appropriate to carry out standard regression analyses with clustered data. As for the adjusted individual level univariate tests, the variance of the intervention effect and other regression coefficients needs to be increased by the design effect

$$1 + (n-1)\hat{\rho}_x\hat{\rho}_y \quad (86)$$

where n is the cluster size, $\hat{\rho}_x$ is the estimate of the intraclass correlation of the covariate or factor, x, and $\hat{\rho}_y$ is the estimate of the intraclass correlation of the outcome, y.[119–121] In univariate tests of intervention effects, this formula simplifies to equation (17): ρ_x is 1 as all individuals in a given cluster are in the same intervention group.

Previously, standard regression methods have been adapted to account for clustering either by including indicator variables to represent the clusters as fixed effects in individual level analyses or by aggregating data to the cluster level. Using the latter approach, analyses are implemented on the cluster-specific summary measures with weights incorporated when cluster sizes are unequal. Both these approaches have well documented shortcomings.[81,122–125]

The use of indicator variables to represent clusters as fixed effects in individual level analyses has four major disadvantages. Firstly, because the intervention effect is confounded with the natural variability between clusters it is not possible to obtain an unbiased estimate of intervention. This is the same confounding which would occur if indicator variables were not used at all. Furthermore, the fixed effects analyses may be even more misleading because the method does succeed in separating cluster variation from individual level variation, thus leading to an artificially large test statistic for the intervention effect and an even smaller p value.[81] Secondly, cluster level covariates may only be incorporated in individual level analyses by disaggregating the data to individual level. This can lead to reduced estimates of standard errors and potentially biased parameter estimates.[125] Thirdly, because the clusters are treated as a fixed rather than a random selection of clusters from a population, it is not appropriate to generalise findings to the entire population of interest. Fourthly, the practice of using indicator variables becomes more inefficient as the number of clusters increases because each variable uses one degree of freedom.

An approach that is being increasingly used to adapt standard regression methods for clustering is the estimation of **robust standard errors (or robust variance estimates)**. This approach allows the requirement that the observations be independent to be relaxed and leads to corrected standard errors in the presence of clustering. Robust standard errors may be derived using the linear, logistic or other regression commands in the statistical package Stata (see chapter 6).

The approach of aggregating individual level characteristics to cluster level also suffers from a number of limitations. Firstly, adjusting for individual level covariates is not straightforward within the framework of cluster level regression analysis. Secondly, it is not advisable to analyse data aggregated from the individual level to the cluster level because relationships at one level do not necessarily hold at another. The erroneous generalisation of relationships that hold at one level to another is sometimes referred to as ecological fallacy.[126] Analyses carried out at a given level can only be reliably used to make inferences at that level, therefore only very limited conclusions can be drawn from a cluster level regression. Thirdly, the number of clusters is generally small in cluster-based studies, and so cluster level regression analysis will be inefficient.[127]

Standard regression analysis at either individual or cluster level does not yield a unified model which reflects the hierarchical structure of the data in cluster-based studies. Single-level analyses will produce results that vary according to the choice of analytical unit, hence some type of hierarchical analysis is required in which both cluster level and individual level covariates may be included. Previously, nested analysis of variance models have been used, but this approach becomes less efficient as clusters vary in size and as more covariates are added.

A number of approaches to the regression analysis of clustered data have been developed which overcome the disadvantages of standard regression methods. The following section describes types of multilevel data structure that commonly occur in cluster-based evaluations and in the next section two regression methods for the analysis of clustered data will be described: random effects modelling (often referred to as 'multilevel modelling') and marginal modelling using generalised estimating equations (GEEs). Specialist packages that are available for implementing regression models for clustered data, and the results produced by the different packages, using a common data set in which clusters were allocated, are compared later in this chapter.

Data structures encountered in cluster-based evaluations

Data from cluster level evaluations have a multilevel structure. Variables are measured at two or more levels, typically with individuals at level 1, nested within clusters (communities or organisations) at level 2. More complicated designs may lead to several levels of clustering. For example, repeated observations (level 1) may be made on patients (level 2) attending GPs (level 3) within general practices (level 4). Four common types of multilevel data structure that occur in health interventions will be described.

Nested data structures

The nested or hierarchical structure is the simplest multilevel structure encountered in cluster-based studies. The defining features are that lower level units are nested within higher units and all elements nested within the same unit at a given level will also share the same unit of any higher level that exists in the data structure. An example of a nested structure is one in which GPs (level 2) are randomised to groups with measurements made on patients (level 1). Patients sharing the same GP will also share the same general practice (level 3).

Cross-classified data structures

In common with the nested design, the lowest level units of interest in the cross-classified design are nested within higher level units. The feature of the cross-classified design which sets it apart from the nested design is that level 1 units grouped within a given level 2 unit do not necessarily also share the same unit at level 3 or at any other higher level. For this reason clustering variables are said to be **crossed** with each other. A cross-classified structure by definition has at least three levels of data. An example of such a structure is one in which the outcome of intervention is studied in relation to area of residence and hospital of treatment. Because of cross-boundary flows there will be cross-classification of area of residence with hospital of treatment.

Repeated measures data structures

In cluster level evaluations, repeated or replicated measurements, rather than the subjects themselves, may be perceived as the lowest level units. The simplest repeated measures structure in a cluster-based study will thus have at least three levels: measurement, individual and cluster. Such data may arise in interventions for which several measures are made of the outcome to check its reliability. Measurements (level 1) are nested within individuals (level 2) who are in clusters (level 3).

Multivariate data structures

Again the measurements themselves are the lowest level units, but there is more than one dependent variable. For example, outcomes may be represented by more than one response variable, as when blood pressure is represented by systolic and diastolic pressures, or when a health status measure gives a profile of scores.

These different multilevel data structures may occur as a result of the study design, the method of sampling used or the nature of the observations themselves.

Random effects models

The random effects model is an extension of the generalised linear model which can be used to allow for the clustering of data within areas or organisations.[120,123,128–130] The term 'random effects modelling' refers to the fact that the mean response is treated as varying randomly between clusters. Several other terms, for example 'multilevel modelling'[120] and 'hierarchical linear modelling',[128] are commonly used to describe the approach and these terms emphasise the multilevel or hierarchical nature of the data being analysed. Other terms include 'variance components

modelling',[122,129] 'mixed modelling',[69] 'random coefficient modelling'[129] and 'contextual modelling'.[124] Many of the advances in random effects modelling have been made in the field of educational research,[122] where pupils are nested within classes and classes within schools. A full discussion of random effects models is provided by Goldstein.[120]

For analysis of a normally distributed outcome on a single covariate or factor, the form of a random effects model is

$$y_{ij} = \alpha + \beta x_{ij} + u_j + e_{ij} \qquad (87)$$

where y_{ij} is the outcome response of the ith individual in the jth cluster, α is the intercept or constant, β is the regression coefficient describing the relationship between the outcome and the covariate or factor, x_{ij} is the covariate or factor value for the ith individual in the jth cluster, u_j is the random effect (level 2) for the jth cluster and e_{ij} is the (level 1) residual for the ith individual in the jth cluster.

It is usually assumed that the u_j are independent and identically distributed (Gaussian) with zero mean and constant variance, σ_u^2 and the e_{ij} in each cluster are independent and identically distributed (Gaussian) with zero mean and constant variance, σ_e^2.

The random effect, u_j, represents the amount by which the intercept for the jth cluster differs from the overall mean value α. The dependence between observations within the same cluster is modelled explicitly via the random effect u_j. It is the presence of the two variance terms u_j and e_{ij} that defines the model as a multilevel or random effects model. The fixed (non-random) part of the model, $y_{ij} = \alpha + \beta x_{ij}$ describes the overall relationship between x and y. When there is no variation between clusters the estimates of the fixed parameters α and β in the random effects model will be the same as those obtained using standard regression analysis.

The simplest random effects regression model is one in which only the intercept varies between clusters (as in equation (87)); this is referred to as a variance components model. Random effects models may be generalised to allow any number of regression coefficients to vary randomly between clusters, and these are referred to as random coefficients or random slopes models. The coefficients will not be equivalent to those which would be obtained by running a separate regression for each cluster because random effects models use information from all clusters in estimating the regression coefficients.[130] The overall regression coefficient estimates are averages of the estimates for each cluster, weighted in such a manner that smaller clusters with higher standard errors contribute less than larger clusters. The cluster-specific coefficients are all **shrunk**[120,122] towards the coefficient of the overall model.

A consequence of treating clusters as random effects is that the results can be generalised to the entire population of clusters, provided the study clusters have been randomly sampled and allocated. Strictly, when clusters have not been randomly selected, inferences only apply to the study clusters. However, it is generally preferable to treat the cluster effect as random rather than as fixed regardless of whether or not allocation is random;[123] in particular, unbiased estimates of the intervention effect cannot be estimated if differences between clusters are treated as fixed effects. In circumstances where there are very few clusters and moderately large numbers of individuals in each, it may be best to implement a univariate cluster level analysis, as there are not enough clusters to estimate the cluster level variation, σ_u^2, efficiently.

Use of random effects regression to estimate components of variance

Random effects modelling can be used to estimate the relative contributions of individual and cluster level characteristics to the variation in continuous outcomes. The terms σ_u^2 and σ_e^2 are equivalent to the between-cluster and within-cluster variance components used to estimate the intraclass correlation coefficient. Most random effects modelling packages provide estimates of the variance components in the output. For binary outcomes the variance components estimates cannot be used to calculate ρ because they are estimated on different scales; the level 1 variation is binomial whilst for other higher levels it is continuous.

Sample size requirements

It is important to emphasise that the sample size requirements for random effects modelling are substantial. While it is not advisable to specify precise numerical requirements, Duncan and colleagues follow Paterson and Goldstein[131] in suggesting that a minimum of 25 individuals in each of 25 clusters may be required for analysis with a single level of clustering.[132] The requirement for a reasonably large number of clusters may limit the application of multilevel approaches to many intervention studies in health care, but application to the analysis of observational evaluations of health

service performance will be appropriate. When the number of clusters is small, variances may be underestimated and confidence intervals may be too narrow, a problem that may be overcome to some extent by adopting a Bayesian approach.[132]

Random effects models for non-continuous outcomes

The estimation of random effects generalised linear models for non-normally distributed outcomes (e.g. random effects logistic regression) presents considerable computational difficulties. Because full maximum-likelihood estimation procedures are not available, approximations (e.g. penalised quasi-likelihood[133]) are used. These may give rise to biased estimates and practical difficulties, for example non-convergence. This is an area of active development.[134–140] Highly computationally intensive Bayesian methods may again be helpful, but their implementation requires a substantial level of expertise (http://www.mrc-bsu.cam.ac.uk/bugs).

Marginal modelling using GEEs

The method of GEEs uses an alternative approach to the regression analysis of clustered data.[141–144] This approach is sometimes referred to as population-averaged or marginal modelling as opposed to the cluster-specific approach used by the random effects model. Whereas in random effects models the cluster level variation is modelled explicitly using u_j, the GEE method treats the dependence between observations as a nuisance parameter. The GEE method assumes a correlation matrix which describes the nature of the association within clusters. The correlation structure is usually assumed to be exchangeable in the context of cluster-based studies, that is, responses from the same cluster are assumed to be equally correlated. Two types of standard error can be obtained under the GEE framework: model-dependent and robust estimates. The latter are consistent even if the within-cluster correlation structure has been specified incorrectly.

In the context of organisation-based evaluations the GEE method suffers from the disadvantages that it requires a large number of clusters[143] and it can only be used to model data with one level of clustering. In spite of these disadvantages the GEE approach is of interest because the regression estimates relate directly to relationships within the overall population rather than specific clusters and because, unlike random effects modelling, correct significance tests may be provided without the need for complex modelling of variance structures.

As with random effects modelling, GEEs produce a solution identical to standard generalised linear models in the absence of clustering. For linear regression the random effects and GEE models will yield identical results in the presence of clustering. For other types of data, such as logistic and Poisson regression, results obtained using the two approaches may differ.

Comparison of specialist random effects modelling packages

Several software packages specifically designed to implement random effects models are available, and include MLn/MLwiN®, HLM®, VARCL®, MIXOR/MIXREG®, MLA® and BUGS®. The last is used to implement Bayesian models. Additionally there are modules within some general statistical packages that permit random effects modelling, including EGRET®, Stata, SAS and BMDP. The programs vary with respect to the type of estimation method and algorithm used but all carry out random effects regression analyses. The algorithmic approaches to fitting random effects models include expectation maximisation, iterative generalised least squares and Fisher scoring for maximum likelihood estimates and the use of Gibbs sampling to produce Bayesian estimates. For maximum likelihood estimation it is recommended that **restricted estimation** be used because it produces unbiased estimates of the random parameters.[120] For full descriptions of these algorithms and others used to implement random effects models, the literature should be consulted.[120,128,129] The general statistical packages Stata®, GENSTAT® and SAS can be used to implement GEEs.

In this section the characteristics and features of the specialist packages MLn, MLA, HLM, VARCL and MIXREG/MIXOR will be described. Some general features of these packages are described in *Table 5*. The reader should also refer to the review by Kreft and colleagues,[145] which described 5V(BMDP)®, GENMOD(SAS)®, HLM, ML3® and VARCL. It is intended here to give a general guide to two of the newer programs and provide updated information on those reviewed by Kreft and colleagues.

MLn (H Goldstein, J Rashbash and M Yang)

MLn supports the analysis of hierarchically structured data with as many as 15 levels of nesting. Regression coefficients may vary randomly at any level. Linear, logistic and Poisson regression models may be implemented as well as log-linear and survival analyses. The default output includes fixed parameter estimates and standard errors, variance estimates with standard errors and significance tests for the fixed and random parameters.

TABLE 5 *General features of the random effects modelling packages*

General features	MLN	MLA	VARCL	HLM	MIXREG
Estimation method	FML/REML	FML/REML	FML	FML/REML	FML
Algorithm for continuous outcome	IGLS	EM	Fisher scoring	EM—Fisher scoring	Fisher scoring
Maximum number of levels	15	2	3[a]	3	2
Number of parameters (fixed and random)	150	128	96	No limit	No limit
Data format	ASCII	ASCII	ASCII	ASCII[b]	ASCII
Data manipulation	Good	None	None	Little	None
Logistic regression	Yes	No	Yes	Yes	Yes[c]
Algorithm for logistic regression	MQL and PQL	–	MQL	PQL	MML
Microsoft Windows®/DOS®	Both	DOS only	DOS only	Both	Both
Batch/interactive	Both	Batch only	Interactive[d]	Both	Both
Documentation/manual available	Yes	Yes	Yes	Yes	Yes

FML, full maximum likelihood; REML, restricted maximum likelihood; IGLS, iterative generalised least squares; EM, expectation maximisation; MQL, marginal quasi-likelihood; PQL, penalised quasi-likelihood; MML, marginal maximum likelihood

[a] *Nine levels may be declared for variance components models where only the constant varies between clusters*
[b] *Data may also be converted directly from a variety of general statistical packages*
[c] *Using MIXOR*
[d] *Interactive for model declaration but the data is imported and prepared in batch mode*

The worksheets together with current model specification can be saved. Other useful features include an option for the analysis of cross-classified models and a command for listwise deletion of cases.

MLn has two estimation procedures for logistic regression: MQL and PQL. The essential difference between the MQL and PQL estimating equations is that the former does not incorporate random effects into the estimation process. Evidence suggests that the MQL procedure is biased, especially in the presence of a large clustering effect,[146] and is essentially equivalent to marginal modelling or population-averaged approaches (see earlier) in which the within-group correlation structure is not explicitly modelled.[133,147] The PQL method, introduced with the current version of MLn, produces estimates that are less biased. Goldstein and Rasbash, however, warn that PQL with a second-order Taylor expansion can produce estimates that are not as efficient as those of a first order Taylor expansion.[148] It is therefore recommended that PQL should be used with a first-order expansion, the analysis should be repeated with a second-order expansion and then the difference between the solutions should be assessed. Even so, Goldstein and Rasbash found that potentially the improvement in using a second-order Taylor expansion for PQL rather than first-order PQL, can be far greater than that in going from MQL to first-order PQL. A Microsoft Windows version of the package, MLwiN, is now available, and facilitates the implementation of Bayesian models.

Web site: http://www.ioe.ac.uk/multilevel/

MLA (F Busing, R van der Leeden and E Meijer)

MLA permits random effects modelling of continuous outcomes only. The default output includes estimates of fixed parameters, their standard errors and p values; random parameters and their standard errors, the intraclass correlation coefficient and the model deviance. The program and manual can be downloaded from the web site.

VARCL (N Longford)

VARCL consists of two separate subprograms: VARL3, which permits up to three-level structures, and VARL9, for up to nine-level structures. VARL9 is for simple variance structures in which the intercept is the only coefficient that varies between clusters, that is, variance component models. Both programs run in batch mode for the data preparation, and are interactive during the modelling stage. A separate data set containing cluster level variables is required for each level of clustering that contributes covariates.

VARCL facilitates the implementation of linear, logistic, Poisson and gamma regression models. The standard output contains fixed and random parameter estimates with standard errors and the model deviance. The worksheet, complete with model specification, can be saved in a dump file for linear regression analyses only.

Web site: http://www.gamma.rug.nl

HLM (AS Bryk, S Raudenbush and RT Congdon)

HLM can be used to implement random effects analogues of linear, logistic and Poisson regression. By default the output contains fixed parameter estimates with standard errors and *p* values, random parameter estimates and a chi-squared test of significance for the between-cluster variance component. For logistic regression, HLM uses the PQL algorithm.

HLM requires a separate data set for each level of clustering regardless of whether the level contributes covariates to the fitted model. HLM allows missing data declaration for individual level data only. A very useful feature of HLM is that as well as importing ASCII data it can also convert data from the statistical packages BMDP, Matlab®, SAS, SPSS®, Stata and SYSTAT®.

Web site: http://www.gamma.rug.nl

MIXREG (D Hedeker and RD Gibbons)

MIXREG can be used to implement random effects linear regression. A separate program, MIXOR, is used for logistic and ordinal regressions, and survival analysis is carried out in MIXGSUR. The standard output contains fixed-parameter estimates with standard errors and *p* values, random parameter estimates and standard errors, and the intraclass correlation coefficient estimate.

Comparison of solutions between regression methods for clustered data

In this section the solutions of several multilevel analysis programs are compared. The data were obtained from an audit of diabetic care at nine primary care health centres with a total of 415 patients. At the five clinics which comprised the intervention group there were specially organised clinics for patients with diabetes while at the other four clinics, diabetic patients were seen in general clinics. *Table 6* illustrates the cluster-based statistics from these data.

Linear regression was used to determine whether there was an association between the natural log of blood glucose concentration and clinic type. Logistic regression was used to find out if there was an association between the proportions of patients receiving dietary advice in the ordinary and specially organised clinics. Several random effects programs were compared as well as GEEs (Stata). Three explanatory variables were used in the linear regressions: clinic type (general or diabetic), sex, and age as a continuous variable. Only clinic type was used in the logistic regressions. Patient data were considered to be clustered within clinics.

Table 7 summarises results of the linear regression analyses carried out in the random effects modelling packages MLn, MLA, VARCL, HLM) and MIXREG. The solution given by the random effects program in Stata (using the command

TABLE 6 *Summary statistics for diabetes data by clinic*

Clinic number	Group number	Cluster size	Mean log of blood glucose level (mmol/l)	Proportion receiving dietary advice
1	2	70	2.34	0.44
2	2	118	2.20	0.23
3	2	61	2.37	0.48
4	2	29	2.21	0.45
5	1	12	2.16	0.58
6	2	64	2.23	0.06
7	1	10	2.19	0.00
8	1	35	2.20	0.31
9	1	16	2.04	0.31

The weighted mean cluster size n_0 *= 43.08*
Intraclass correlation coefficient for log of blood glucose level = 0.0197
Intraclass correlation coefficient for dietary advice = 0.113

TABLE 7 *Comparison of random effects model results (for explanation see text)*

	OLS	Stata	MLn	MLA	VARCL	HLM	MIXREG
Coefficient of clinic type	0.115	0.123	0.123	0.123	0.123	0.123	0.123
Standard error	0.055	0.0636	0.063	0.063	0.063	0.063	0.063
t ratio	2.089	1.94	1.94	1.94	1.94	1.94	1.94
p value	0.037	0.053	0.052	0.052	0.052	0.052	0.052
Coefficient of age	–0.004	–0.004	–0.004	–0.004	–0.004	–0.004	–0.004
Coefficient of sex	0.065	0.065	0.065	0.065	0.065	0.065	0.065
Constant	2.339	2.342	2.342	2.342	2.342	2.342	2.342
Between-cluster variance	–	0.00197	0.00197	0.00197	0.00197	0.00197	0.00197
Within-cluster variance	–	0.1786	0.1786	0.1786	0.1786	0.1786	0.1786
Intraclass correlation	–	0.0110	0.0110	0.0110	0.0110	0.0110	0.0110

OLS, ordinary least squares

'xtreg') is also given. Standard linear regression (using ordinary least squares), which takes no account of clustering, suggests that the relationship between blood glucose control and clinic type is significant, with a *p* value of 0.037. A positive coefficient for clinic type suggests that patients in the special clinics have a higher blood glucose level than those in the general clinics.

Each of the specialist random effects packages gives the same result, which differs from that obtained using standard linear regression. The consequence of allowing for clustering is a larger standard error and a higher *p* value, and this confirms that an incorrect single level analysis could give misleading results. The Stata command 'xtreg' also gives the same result. However, it is important to be aware that there are several options for the 'xtreg' command, and it is the maximum likelihood option 'mle' which gives results which correspond to MLn. The package Stata may be more convenient for general use, but it is not able to handle the range of complex data structures accommodated by MLn.

Table 8 summarises the results of logistic regression analyses. The packages used were MLn, VARCL, HLM and MIXOR, and Stata for GEEs with and without robust estimates of variance using the 'xtgee' command. The MLn solution was estimated using the PQL algorithm with second-order Taylor expansion.

The result obtained using standard logistic regression differs from that obtained using methods that allow for clustering. There are clear differences between the results obtained using GEEs and those obtained using the random effects packages, and the results obtained using the random effects packages are less consistent than those obtained for the continuous outcome. Goldstein[149] and Rodriguez and Goldman[146] noted the similarity between the non-linear procedures employed by MLn and VARCL and this is borne out by the results for the logistic regression. However, besides these two programs it appears that the solutions for the logistic regression generally differ among the packages. This might be expected as the extension of adjusted regression methods for dichotomous outcomes is still in the developmental stage.[113] The estimation procedures used for logistic regression differ, and they can produce differing estimates. De Leeuw and Kreft[150] stress that there is no uniformly best method for

TABLE 8 *Comparison of logistic regression results*

	Standard logistic regression	MLN using second-order PQL	VARCL	HLM	MIXOR	GEE model-dependent variance estimates	GEE robust variance estimates
Log odds ratio of clinic type	–0.05135	0.0784	0.078827	0.032104	0.06433	0.090	0.090
Standard error of log odds	0.278020	0.5357	0.538199	0.636890	0.71374	0.579	0.572
Odds ratio of clinic type	0.949945	1.081555	1.082017	1.032625	1.066444	1.095	1.095

logistic regression. Analysis of simulated data with known population parameters is the ideal approach for gauging the bias in estimates incurred by these programs. For the purposes of this review, we use the diabetes data to draw attention to the fact that there are differences between the estimation procedures used by different packages and to observe that different results may be obtained by different methods.

Applications of regression methods for clustered data in health and healthcare research

The practical application of regression methods for clustered data to the evaluation of health care is still at an early stage of development. Some applications up to 1996 were reviewed by Rice and Leyland,[123] and up to 1998 by Duncan and colleagues.[132]

When regression models are used, methods appropriate for clustered data should be employed in order to obtain the correct standard errors and confidence intervals. GEEs may be used to adjust for the clustering of individual responses at organisation or area level. This approach may be applicable if cluster sampling is used to recruit individuals from different clinics.[151] The approach will also be relevant when cluster level covariates are to be included in analyses. For example, in a study of neonatal mortality in relation to activity level in neonatal units, GEEs were used to adjust for the correlation of infant outcomes within hospitals.[152]

Duncan and colleagues[132] make the point that random effects models will be more useful when the evaluator wants to distinguish the effects of the organisational or geographical context from the composition of the sample of individuals within the organisation or area. This type of distinction is often important in observational evaluations of existing health services, for example, in comparing the performance of different institutions, or comparing healthcare processes and outcomes in different geographical areas. Goldstein and Speigelhalter[153] discussed some of the statistical issues in comparisons of institutional performance. They showed that it is important to carry out the correct hierarchical analysis which models the nesting of individual responses within hospitals or health authorities. This type of analysis in which organisational units are treated as random effects leads to point estimates which are shrunk towards the population mean with confidence intervals that are more precise when compared with a conventional fixed effects analysis. However, when ranking of outcomes by organisation is important, confidence intervals for ranks may be so large that they are not informative. Examples of the application of multilevel analysis to geographical variations in health are described by Duncan and colleagues.[132]

An example: analysis allowing for clustering

To illustrate how an analysis might be performed in order to allow for clustering, we analysed data obtained from an audit of clinical care (*Table 9*). The aim of this analysis was to see whether there was a difference between the proportion of attenders who were receiving antihypertensive treatment at health centre clinics compared with attenders at single-handed GPs.

TABLE 9 *Data from clinical audit. Figures are frequencies (percentage of row total)*

Clinic type	Number of patients	Number on blood pressure treatment	Mean age (years)	Number of women (%)
HC	62	47 (76)	63	46 (74)
HC	51	44 (86)	62	34 (67)
HC	43	41 (95)	65	33 (77)
HC	43	32 (74)	64	27 (63)
HC	35	28 (80)	63	26 (74)
HC	21	9 (43)	69	7 (33)
HC	83	69 (83)	66	61 (73)
HC	68	24 (35)	56	43 (63)
HC	59	28 (47)	60	34 (58)
HC	74	39 (53)	59	45 (61)
HC	67	53 (79)	64	52 (78)
HC	44	27 (61)	59	28 (64)
HC	63	44 (70)	62	40 (63)
HC	76	39 (51)	63	49 (64)
HC	36	15 (42)	55	27 (75)
HC	66	41 (62)	60	42 (64)
HC	37	24 (65)	64	27 (73)
HC	47	32 (68)	57	42 (89)
GP	45	5 (11)	52	28 (62)
GP	43	6 (14)	52	29 (67)
GP	36	9 (25)	54	31 (86)
GP	89	14 (16)	46	69 (78)
GP	120	19 (16)	46	81 (68)
GP	21	8 (38)	53	13 (62)
GP	120	31 (26)	54	87 (73)
GP	84	19 (23)	47	49 (58)
HC sub-total	975	636 (65)	61	663 (68)
GP sub-total	558	111 (20)	49	387 (69)
Total	1533	747 (49)	57	1050 (68)

HC, health centre

Initial inspection of the data showed that there were 26 clinics including 18 health centres and eight single-handed GPs. The number of patients sampled per clinic was quite variable. The proportion of patients treated for hypertension seemed to be higher at the health centre clinics but the health centre patients were older than those attending the GPs. The value for ρ estimated from the 26 clinics was 0.259, but it might not be justified to assume a common value for ρ at both types of clinic, since the estimated value for ρ in the health centre clinics was 0.107 while for the GP clinics it was 0.013. However, the number of GP clinics was small and possibly not sufficient to obtain a stable estimate for ρ.

These observations suggest that certain approaches to analysis will be more appropriate than others. Because age is associated with both blood pressure and clinic type, it will be necessary to carry out analyses which allow for differences in age among clinic types. Donner and Klar[113] pointed out that adjusted individual level hypothesis tests which assume a common value for the design effect across groups will be less applicable to non-randomised data.

The results of several different methods of analysis are shown in *Table 10*. Initially, cluster level analyses were performed using the cluster-specific proportions as observations. A non-parametric rank sum test shows a significant difference between the two clinical settings with an approximate *z* statistic of 3.95. The two-sample *t* test gave a mean difference in the cluster-specific proportions of 0.439 (95% confidence interval 0.307–0.571). Analyses were also performed at the individual level. An analysis which made no allowance for the clustering of responses gave a difference between the two proportions of 0.453 with 95% confidence intervals from 0.409 to 0.498 and a *z* statistic of 17.1. However, a standard statistical analysis would be incorrect because, as we noted above, between-clinic variation is present. In order to construct confidence intervals, the design effect should be incorporated into the formula for the standard error for the difference in two proportions as shown in equation (55). The design effect for the health centre clinics is 6.66, and for the single-handed GPs is 1.90. The standard error of the difference in proportions was 0.0458. The difference was still 0.453, but the corrected confidence intervals were from 0.363 to 0.543. This result showed that the difference between the two audits was estimated less precisely after allowing for between-clinic variation. The corresponding adjusted chi-squared test[113] yielded a χ^2 of 113.0

TABLE 10 *Results obtained using different methods of analysis to compare the proportion of attenders on antihypertensive treatment at health centre clinics and at single-handed GPs*

Method	Estimate (95% confidence intervals)	*z* statistic
Cluster level analysis	**Difference (95% confidence interval)**	
Rank sum test	–	3.95
Two-sample *t* test	0.439 (0.307–0.571)	$t = 6.85$
Individual analysis – difference of proportions		
Standard univariate test	0.453 (0.409–0.498)	17.1
Univariate test adjusted for design effect	0.453 (0.363–0.543)	$\chi^2 = 113.0$ (df = 1)[a]
Univariate individual analysis – logistic regression methods	**Odds ratio (95% confidence interval)**	
Standard logistic regression		
Univariate	7.56 (5.91–9.66)	16.1
Adjusted for age and sex	5.46 (4.22–7.08)	12.9
Logistic regression, robust standard errors		
Univariate	7.56 (4.88–11.69)	9.1
Adjusted for age and sex	5.46 (3.66–8.15)	8.3
GEEs		
Univariate	7.32 (4.62–11.57)	8.5
Adjusted for age and sex	5.06 (3.32–7.70)	7.6
Random effects logistic regression		
Univariate	8.47 (4.51–15.91)	6.6
Adjusted for age and sex	6.16 (3.39–11.18)	6.0

df, degree of freedom

[a] *Refers to adjusted chi-squared test*[59]

(one degree of freedom). Note, however, that this test assumes a common value for ρ.

An alternative approach to the analysis is to use logistic regression. A conventional logistic regression analysis showed that the relative odds of patients being on antihypertensive medication in the health centre clinics compared with the GP clinics were 7.56 (5.91–9.66). The *z* statistic obtained from this model was 16.1. This analysis was modified to allow for clustering within clinics by estimating robust standard errors giving the same odds ratio 7.56, but the 95% confidence intervals were from 4.88 to 11.69 with a *z* statistic of 9.1. Analyses were adjusted for age and sex and the relative odds were reduced to 5.46 (3.66–8.15). The method of GEEs offered an alternative approach to regression analysis. This was carried out by specifying a logistic link function, an exchangeable correlation structure, and robust standard errors to allow for clustering within clinics. The analysis adjusted for age and sex gave an odds ratio of 5.06 (3.32–7.70) which was similar to the result obtained using logistic regression with robust standard errors, but this analysis adjusted the estimate for clustering as well as the standard error. We used the package Stata to perform both the logistic regression and GEE analyses.[154]

A third approach to the analysis of these data is to use random effects logistic regression to model the variation in response at clinic level. We used a simplified model which assumed the same variation at level 2 (i.e. clinic level) for both settings. The adjusted analysis gave a higher odds ratio and slightly wider confidence intervals, 6.16 (3.39–11.18), when compared with the logistic regression and GEE analyses. The level 2 variance on the log odds scale was 0.387 with a standard error of 0.139. We used the package MLn[155] to fit the random effects logistic regression models with second-order PQL estimation.

The incorrect individual level analyses with no adjustment for clustering gave narrower confidence intervals and larger *z* statistics than any of the methods which allow for clustering of responses. Carrying out a two-sample *t* test using the cluster-specific proportions as observations provided the most accessible method of analysis and gave a result which allowed for between-cluster variation. However, as it was also necessary to adjust for individual level confounders, one of the regression methods for clustered data was to be preferred. Logistic regression with robust standard errors gave a similar result to that obtained using GEEs in these data. Logistic regression was more accessible, and was computationally less intensive with shorter analysis times, but the GEE approach adjusted both the estimate and the standard error for clustering, and was to be preferred for this reason. Random effects logistic regression might have allowed additional modelling of the variance at cluster level, which could have been useful if the analysis was to focus on the performance of particular primary care clinics.

Chapter 7

Twelve methodological recommendations

We summarised the preceding review into 12 methodological recommendations:

(1) Recognise the cluster as the unit of intervention or allocation. Healthcare evaluations often fail to recognise, or correctly utilise, the different levels of intervention which may be used for allocation and analysis.[22] Failure to distinguish individual level from cluster level intervention or analysis can result in studies which are inappropriately designed or which give incorrect results.[109] Researchers should recognise different organisational levels at which the intervention may be made and distinguish between them when deciding on the method of allocation or analysis.

(2) Justify the use of the cluster as the unit of intervention or allocation. For a fixed number of individuals, studies in which clusters are randomised to groups are not as powerful as traditional clinical trials in which individuals are randomised.[109] This is because individual responses within clusters are correlated with each other in cluster-based studies which thus usually require more subjects to obtain comparable power. The decision to allocate at organisation level should therefore be justified on theoretical, practical or economic grounds (*Box 3*).

(3) Include a sufficient number of clusters. Generally, the evaluation of an intervention which is implemented in a single cluster will not give generalisable results. For example, a study evaluating a new way of organising care at one diabetic clinic would be an audit study from which generalisable observations may be difficult. 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 the underlying variation in response between the two clinics. Studies with only a few (less 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.[36] Studies with as few as six clusters per group have been used to demonstrate effects from cluster-based interventions,[156] but larger numbers of clusters may be needed, particularly when the investigator is interested in a relatively small intervention effect. In principle, as many clusters as possible should be included in an evaluation, but in practice the number of clusters may be limited by practical or financial constraints.

BOX 3 Reasons for carrying out cluster-level evaluations

(1) Public health and health care programmes are generally implemented at organisation rather than individual level, so cluster level studies are more appropriate for assessing the effectiveness of such programmes

(2) 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

(3) '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

(4) 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

(4) Randomise clusters wherever possible. While clinical trials are considered essential for the evaluation of treatments aimed at individuals, random allocation of clusters in organisation level interventions has been less common. Randomisation is used to ensure that the estimate of programme effect is not biased as a result of confounding with known or unknown variables. Recent examples of cluster randomised studies include the British Family Heart Study[157] and the Mwanza study of HIV prevention.[156] There are occasions when the investigator will not be able to control the assignment of clusters, for instance when evaluating an existing service or policy.[75] However, because of the risk of bias, use of quasi-experimental or observational designs should always be explicitly justified.

(5) In non-randomised studies include a control group with observations before and after the

intervention. When randomisation is not feasible, a control group should be included. Each group should include a sufficient number of clusters (see point 3) which should be stratified for important prognostic factors so far as possible (see point 8). A wide range of confounders should be measured. Outcome variables should be measured both before and after the intervention.

(6) In single group studies include repeated measurements over time. Sometimes it is not feasible to include a control group, as for example when a new policy is implemented at national level. In this case, repeated assessments should be made both before and after the intervention in order to control for secular changes in the outcome.

(7) Allow for clustering when estimating the required sample size. When the cluster is both the unit of intervention and the unit of evaluation, conventional statistical approaches may be used to estimate the required number of clusters, using cluster level measures of outcome as the units of observation. However, when evaluating cluster level interventions by means of observations made at the individual level, standard sample size formulae will not be appropriate for obtaining the total number of individuals required. This is because they assume that the responses of individuals within clusters are independent.[19,60,86,108,109,111] Individuals within clusters are usually more similar than those in different clusters. Dependence between subjects within clusters needs to be recognised when estimating the sample size. Standard sample size formulae underestimate the number of individuals required because although variation **within clusters** is allowed for, variation **between clusters** is not.

Methods for calculating the total number of individuals required for cluster level interventions typically involve adjusting standard formulae to allow for the correlation between subjects.[109] A quantity known as the design effect or variance inflation factor is used to adjust standard sample size formulae, in order to 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

$$\text{Deff} = 1 + (n - 1)\rho$$

where Deff is the design effect, n is the average size of the clusters and ρ is the intraclass correlation coefficient for the outcome of interest.

The design effect may be interpreted as the number of times more subjects a cluster-based evaluation should have, compared with one in which individuals are randomised, in order to attain the same power. ρ is the proportion of the total variation of the outcome that is between clusters; it essentially gauges the degree of similarity or correlation between subjects within the same cluster. The larger ρ is, that is, the more similar the subjects are within a cluster, the greater the size of the design effect and the larger the additional number of subjects required in an organisation-based evaluation to compensate for the loss in power. In studies of large organisational units, ρ usually takes small positive values; however, with large numbers of individuals per cluster even a small ρ can lead to a large design effect.

For different studies with the same outcome, the estimate of ρ is more comparable than the design effect because it is not dependent upon the number of subjects selected from within each of the clusters. It is therefore ρ that should be known or estimated prior to the study. If ρ is not available plausible values must be guessed. One of the recommendations of this review is that researchers should publish estimates of ρ for key outcomes of interest.[103] This will aid the planning of future organisation level interventions. Some examples of values for ρ are given in chapter 9.

The investigator will often find that clusters are of predetermined size. The number of clusters required can then be estimated by dividing the total number of individuals required by the average cluster size. When it is feasible to sample individuals within clusters, the power of the study may be increased either by increasing the number of clusters or the number of individuals within clusters.

Increasing the number of clusters rather than the number of individuals within clusters has several advantages.[61] The result of the study will usually appear to be more generalisable if the intervention has been implemented in a number of different clusters. A larger number of clusters also allows more precise estimation of ρ and more flexible approaches to analysis.[61] Furthermore, there is a limit to the extent to which power may be increased solely by increasing the number of individuals within clusters.[61] However, the relative cost of increasing the number of clusters in the study, rather than the number of individuals within clusters, will be an important consideration when deciding on the final structure of the sample.

Appropriate formulae for sample size calculations are given by Cornfield,[109] Donner and colleagues,[86,108,111] Hsieh[19] and Shipley.[60]

(8) Consider the use of matching or stratification of clusters where appropriate. In cluster-based studies, randomisation may be stratified to ensure that treatment groups are balanced with respect to important prognostic factors, so as to increase the power of the study. Stratification is of particular importance when the number of clusters randomised is small (as is typically the case in organisation level evaluations) since randomisation will not ensure that treatment groups are balanced. 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 cluster-based trials where there are known prognostic factors, unless the number of clusters is quite large. Stratification may increase power in randomised studies and reduce bias in quasi-experimental studies; however, it is only useful if the prognostic factor(s) is(are) 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. This design might seem preferable when the number of prognostic factors is large relative to the number of study clusters.[50] 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.[51] Secondly, when the number of clusters is less than about 20, the loss of degrees of freedom associated with a matched analysis may result in serious loss of statistical power,[55] although this will depend on the strength of the association between the matching variable and the outcome variable. It will often be more advisable to include at least four clusters in each stratum. If matching is felt to be essential, it is worth considering the use of an unmatched cluster level analysis to evaluate the intervention effect.[34]

(9) Consider different approaches to repeated assessments in prospective evaluations. There is a choice between two basic sampling designs for follow-up studies in which the organisation is the unit of intervention: 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 new sample of subjects is drawn from the clusters at each measurement occasion.[67,69,71] The cohort design is most appropriate when the focus of the study is on the effect of the programme at the level of the subject,[67] as one can link changes in outcome to individual level prognostic factors. The repeated cross-sectional design on the other hand is more appropriate when the focus of interest is the long-term effect of intervention on some cluster level index of health such as disease prevalence. This is because the repeated cross-sectional design is more likely to be representative of the clusters at the later measurement occasions, particularly for studies with long follow-up. Choice of design should be dictated by the questions and hypotheses driving the research. For studies where either design is theoretically appropriate, 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 this acts to reduce the variation of the estimated intervention effect.[69]

When repeated cross-sectional sampling is used, it may be important to anticipate the possible direction and size of secular trends in the outcome, as the secular trend may compromise the power of the study to detect an intervention effect. In healthcare settings it must be remembered that secular changes in case mix may be substantial.

(10) Allow for clustering at the time of analysis. Standard statistical methods are not appropriate for the analysis of individual level data from cluster-based evaluations, because they assume that the responses of different subjects are independent.[109] 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.

Univariate tests of intervention effect for organisation level evaluations with individual level outcomes, may be done in three ways: (a) analysis at the level of the cluster, applying standard statistical methods to the cluster means, proportions or log odds; (b) analysis at the level of the individual using formulae for which the variance of the estimate has been adjusted to allow for the similarity between individuals; and (c) analysis at the level of the individual using regression methods for clustered data to implement univariate tests of the intervention effect.

The choice of level of analysis is sometimes referred to as the 'unit of analysis problem'.[22] Cluster level analyses have the appeal that that the unit of intervention is the same as the unit of evaluation, but individual level analysis may seem more appropriate on theoretical grounds when the aim of

cluster level intervention is to modify individual level outcomes. For example, a health promotion intervention might be designed to modify the coronary risk factors of individuals in the study population.

Individual level analyses must be adapted to allow for dependence between individuals within the same cluster. This is done by incorporating the design effect into standard formulae. For adjusted individual level analyses, ρ is estimated from the study data and a weighted form of the mean cluster size is usually used to calculate the design effect. About 20 clusters are required to estimate ρ with a reasonable level of precision. Appropriate formulae are referenced by Donner and Klar[11] for adjustments to methods used for univariate tests at the individual level. Methods for estimating confidence intervals are given by Donner and Klar.[59]

The adjustments to standard univariate methods for application at the individual level only adjust the standard error of the intervention effect for clustering; they do not adjust the intervention effect itself. For this reason, regression methods for clustered data might be expected to provide estimates that are more efficient and, thus, provide correct significance tests.

(11) Allow for confounding at both individual and cluster levels. When there is a need to adjust the estimate of intervention effect for individual and/or cluster level prognostic variables, then statistical methods which allow for similarity between individuals in the same cluster should be used. Regression methods for clustered data such as random effects modelling and marginal modelling using GEEs allow for correlation between subjects and both individual level and cluster level prognostic factors can be included in the analysis.[123,141] The techniques can be used for studies with clusters that vary in size.

A variety of specialist computer packages for estimating random effects models are available.[145] In the UK the most popular is MLn. The general statistical packages Stata[154] and SAS[100] can be used to implement GEEs. Application of random effects models is more appropriate when the number of clusters studied is large enough to estimate between-cluster variation (around 20) and a similar number of clusters is required for the use of GEEs.

In addition to linear models some of the random effects modelling packages also provide logistic and Poisson regression and other generalised linear models. In general, the estimation procedures used for such models are approximate, and have been reported to produce biased estimates in some circumstances. This is a rapidly developing field, and results from random effects models with non-normal errors must be treated with some caution.

(12) Include estimates of intracluster correlation and components of variance in published reports. For reasons presented earlier, estimates of ρ are required for sample size calculation at the design stage of organisation-based evaluations. However, there is a danger in extrapolating ρ from one study to another because one of the components of variance (either the within-cluster or between-cluster component) may change from one population to another, or depend on the sampling strategy. The between- and within-cluster components of variance should therefore be reported in addition to ρ.

Concluding remarks

The main guidelines that have arisen from this systematic review of organisation level evaluations are summarised in *Box 4*. Investigators will need to consider the special circumstances of their own evaluation and use discretion in applying these guidelines to specific circumstances. We also emphasise that the conduct of cluster-based evaluations may present special difficulties. The issue of informed consent needs careful consideration (see chapter 3). Interventions and data management within clusters need careful definition and standardisation. 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.

BOX 4 Checklist for design and analysis of area and organisation-based interventions

(1) Recognise areas or organisational clusters as the units of intervention

(2) Justify allocation of entire clusters of individuals to groups

(3) Include a sufficient number of clusters. Studies in which there are less than four clusters per group are unlikely to yield conclusive results

(4) Randomise clusters to intervention and control groups whenever possible, and justify use of non-randomised designs

continued

BOX 4 *contd* Checklist for design and analysis of area and organisation-based interventions

(5) In non-randomised designs include a control group and measure outcome variables before and after the intervention

(6) When only a single group can be studied, include repeated outcome measurements before and after the intervention

(7) Multiply standard sample size formulae by the design effect in order to obtain the number of individuals required to give a study with the same power as one in which individuals are randomised. Estimates of the intraclass correlation coefficient should be obtained from earlier studies

(8) Consider stratification of clusters in order to reduce error in randomised studies and bias in quasi-experimental studies. Some stratification should usually be used unless the number of clusters is quite large. Researchers should be aware of the limitations of the matched-pairs design (i.e. a design with only two clusters per stratum)

(9) Choose between cohort and repeated cross-sectional sampling for studies that involve follow-up. The cohort design is more applicable to individual level outcomes, and may give more precise results but is more susceptible to bias. The repeated cross-sectional design is more appropriate when outcomes will be aggregated to cluster-level, and is usually less powerful, but is less susceptible to bias

(10) Standard statistical methods, applied at the individual level, are not appropriate because individual values are correlated within clusters. Univariate analysis may be performed either using the cluster means or proportions as observations, or using individual level tests in which the standard error is adjusted for the design effect. Where there are fewer than about 10 clusters per group, a cluster level analysis may be more appropriate

(11) When individual and cluster level prognostic variables need to be allowed for, regression methods for clustered data are appropriate. Provided there are sufficient clusters, use of regression methods for clustered data may also provide a more flexible and efficient approach to univariate analysis

(12) Authors should publish estimates of components of variance and the intraclass correlation coefficient for the outcome of interest when reporting organisation level evaluations

Chapter 8

Case study: review of publications in seven health science journals

- **Objectives.** To identify the main departures from good practice in evaluations of area- or organisation-based evaluations.
- **Methods.** A survey of seven peer-reviewed health science journals identified 56 papers which reported evaluations of area or organisation-based interventions which were reviewed to identify the main departures from recommendations.
- **Results.** Few studies explicitly considered the distinction between clusters and individuals as levels of intervention and evaluation. Intervention studies were often implemented in small numbers of clusters, sometimes without a control group, and often without randomisation. Analysis at the individual level did not usually include adjustment for correlation of outcomes within clusters.
- **Conclusions.** There is a need to recognise the different levels of organisational clustering at which interventions may be implemented, to include sufficient clusters in intervention and control groups, and to allow for correlation of outcomes within clusters when analysis is at the individual level.

Introduction

Health interventions have traditionally been evaluated using the approaches of clinical epidemiology in which the individual subject is regarded as the unit of intervention and analysis. Health interventions are commonly implemented, not for individual subjects, but for entire clusters of individuals in geographical areas or units of health service organisation such as health authorities, hospitals or general practices. Examples of interventions implemented at these levels include screening programmes, medical practice guidelines and health promotion interventions.

The evaluation of health interventions which are implemented at cluster level presents several problems (see chapter 1). Firstly, outcomes may be evaluated either at cluster level or at individual level. It is important to distinguish between the different levels at which evaluation may be achieved. Secondly, it may only be possible to include a small number of clusters in a study. For example, only a small number of hospitals or general practices may be available for investigation. Thirdly, the responses of individuals within geographical or organisational clusters tend to be more similar to each other than to those in other clusters. The correlation of individual responses within clusters means that between-cluster variation must be allowed for in the design and analysis of cluster level evaluations.[11,109] There is evidence to suggest that the problems associated with evaluating cluster level interventions are not sufficiently widely appreciated nor addressed in a satisfactory manner.[9,12,22]

The aim of this part of the review was to identify the main departures from good practice in area-wide and organisation-based evaluations.

Methods

Study identification

Journals selected for study were the *Journal of Public Health Medicine, Journal of Epidemiology and Community Health, British Medical Journal (BMJ), Journal of the American Medical Association (JAMA), Medical Care,* the *International Journal of Technology Assessment in Health Care* and the *European Journal of Public Health.* These were selected because they represent major European and North American journals which publish papers in the field of healthcare evaluation. We selected for study issues from 1996, the most recent complete year. For the two weekly journals (*BMJ* and *JAMA*) we only included the first issue from each month in 1996, in order to avoid over-representation of papers from these journals. For the other journals we included all issues published in 1996. We handsearched each issue of the journal, and included papers reporting primary research or secondary data analyses which studied interventions implemented at the level of organisation or geographical area.

Data collection

For each study two independent observers (MG and SC) abstracted data on to a standard proforma. Data collected included details of the study design; whether it was an intervention study or evaluation of an existing service; the number of groups for

comparison; the unit of intervention or comparison; whether allocation was by randomisation and, if not, whether a reason was given; whether outcome assessment was after only or before and after; the method of identifying units of observation; and, for before and after evaluations, whether a cross-sectional or cohort design was used. We also noted the number of individual level units in the study; the numbers of levels of clustering that were present; the number and type of clusters at each level; whether a sample size calculation was reported; and whether it made allowance for clustering of the data. We determined the type of outcome variable and the main level of analysis. For individual level analyses, we determined whether allowance was made for clustering; whether standard errors were adjusted for the design effect directly; whether regression analysis was used, and, if so, whether and how it allowed for clustering. We also recorded whether confounding variables were recorded at individual and cluster level and whether they were included in the analyses. Finally, we noted whether the intraclass correlation coefficient was recorded. The quantitative findings of the survey are presented by tabulation of frequencies. Examples were also selected to provide qualitative illustrations of the findings.

Results and discussion

The search process resulted in the identification of 56 reports which fulfilled the eligibility criteria.[47,152,158–211] Their source and characteristics are shown in *Table 11*. The papers reported evaluations of a range of services including health promotion, population screening, and primary and community care as well as hospital-based interventions. The most frequent departures from good practice are summarised in *Box 5*. The following subsections discuss and illustrate some of the methodological problems identified.

TABLE 11 *Characteristics of studies included in this review*

Variable	Frequency (%)
Journal	
Journal of Public Health Medicine	9 (16)
Journal of Epidemiology Community Health	13 (24)
BMJ	13 (24)
JAMA	6 (11)
Medical Care	9 (16)
International Journal of Technology Assessment in Health Care	2 (4)
European Journal of Public Health	4 (7)
Country of origin	
UK	22 (39)
USA	14 (25)
Other western European countries	17 (31)
Australasia	1 (2)
Canada	2 (4)
Disease or problem	
Infectious disease	1 (2)
Cancer	9 (16)
Blood forming organs	2 (4)
Mental health	4 (7)
Cardiovascular disease	4 (7)
Respiratory disease	3 (6)
Musculoskeletal	4 (7)
Genitourinary	1 (2)
Maternal and child health	5 (9)
Injuries	1 (2)
External causes	1 (2)
Factors influencing health status and contact with health services	21 (38)
Type of service	
Health promotion	4 (7)
Population screening	5 (9)
Primary medical care	13 (24)
Community care	4 (7)
Hospital, medical	7 (13)
Hospital, surgical	2 (4)
Hospital, orthopaedic and trauma	2 (4)
Hospital, obstetrics and gynaecology	5 (9)
Hospital, other	7 (13)
Other	7 (13)

Failure to recognise areas or organisational clusters as units of intervention

Many of the studies included in the review did not specifically acknowledge different levels of organisational clustering present in the data, nor the extent to which these might be regarded as levels of intervention or evaluation.

Example: comparison of patient satisfaction with ambulatory visits in competing healthcare delivery settings in Geneva, Switzerland[175]

A survey was carried out including 1027 patients who were sampled from one managed care organisation, one private group practice and one university hospital outpatient clinic. Individual level analyses were carried out to see if patient satisfaction varied in the different care settings. Patients attending the managed care organisation were less satisfied, perhaps because they could not freely choose their doctor.

Comment. The study design suffered from the weakness that only one example of each type of care setting was included in the study. Satisfaction with care was likely to vary among managed care organisations, so more than one organisation should have been represented. Within care settings, satisfaction

BOX 5 Common departures from recommendations

(1) Failure to recognise areas or organisational clusters as units of intervention
(2) Neglect of randomisation
(3) Evaluation of interventions implemented in a single cluster
(4) Lack of control group
(5) Lack of appropriate sample size calculations
(6) Individual level analysis without adjustment for correlation of responses within areas or organisations
(7) Inappropriate methods of analysis used to allow for clustering of responses or to adjust for cluster level covariates in individual level analyses
(8) Cluster level analysis without inclusion of individual level covariates

might vary according to the doctor visited; this second level of clustering could also be recognised in the design and analysis of the study.

Neglect of randomisation

The randomised controlled trial has come to be accepted as the preferred method for evaluating individual level interventions but has been less used in the evaluation of organisation level interventions. In this sample, only four of 56 studies were randomised, and in each case randomisation was at the individual level. Recent reports have demonstrated the feasibility of cluster level randomisation as a method for evaluating the effectiveness of cluster level interventions.[48,156,157] Four reasons why cluster level evaluations should be carried out are shown in *Box 3*.

Some of the studies included in our survey appeared to represent *ad hoc* evaluations of new or existing interventions which had been implemented without planned evaluation in mind. The rationale for selecting a particular level of clustering as the unit of intervention and the reasons for avoiding use of randomisation were not usually discussed. Examples included evaluations of a 'drop-in' service for women in a single district,[160] a smoking prevention programme for schoolchildren in one health region,[158] and helicopter emergency ambulance services.[159]

Example: evaluation of Grampian Smokebusters – a smoking prevention initiative aimed at young teenagers[158]

In 1987 a club for children aged 10–13 years was launched in the Grampian region of Scotland with the aim of discouraging them from smoking. Evaluation was by means of a questionnaire administered to pupils at 27 primary schools and 40 secondary schools. Comparison was made with data from national surveys in Scotland. Comparisons were also made between club members and non-members in the study area.

Comment. The evaluation suffered from the limitation that the intervention was implemented in a single study area. Comparisons with national data would be completely confounded by underlying differences between Grampian and the rest of Scotland. Within the study area, because club membership was not allocated at random, comparisons between club members and non-members were likely to be biased. Correlation of outcomes within schools or classes was not considered in the analysis. A preferred design would be to randomise classes or schools to club membership.[40] Individual level randomisation might be feasible, but contamination is likely to be a problem.

Intervention in a single cluster

Evaluations often described interventions implemented in a single cluster. Examples included models of organisation implemented in a single anticoagulant clinic,[198] or health promotion interventions implemented in a town,[169] health region[158] or country.[203] Where a single cluster is studied, it is clear that the results cannot be considered to be generalisable, but such studies will be of local interest and can be used to inform the development of local services.

Need for a control group

When an intervention is implemented in a single cluster, the size of the intervention effect can only be gauged by before and after evaluation. However, because the outcome may be influenced by factors other than the intervention, a comparison group is also needed. With only one or a few clusters per group it may be difficult to distinguish the intervention effect from the natural variability between clusters. For this reason, the study should include sufficient clusters to allow estimation of both the extent of between-cluster variation and the size of the intervention effect. However, in this survey there were examples where the number of control clusters was too small to gauge the extent of between-cluster variation.[158,169]

Example of a study with intervention in a single cluster without a control group: methods for managing the increased workload in anticoagulant clinics[198]

In order to cope with the increased workload at their anticoagulant clinic the investigators arranged for healthcare assistants to see most patients at their

routine visits, so that doctors could focus their attention on those patients with special requirements. The quality of anticoagulant control was audited before and after the change in clinic organisation.

Comment. Because the intervention was implemented at a single clinic with no comparison group it may be difficult to generalise the results of the study. In addition, because case mix may change over time, it may not be justified to conclude that changes in anticoagulant control were the result of the intervention.

Example of a study with insufficient clusters in intervention and control group: cost-effectiveness and equity of a community-based cardiovascular disease prevention programme in Norsjo, Sweden[169]

A community-based health promotion programme was implemented in Norsjo in northern Sweden. Comparison was made with two reference counties, Norrbotten and Vasterbotten.

Comment. The intervention was implemented in a single area. Two control areas were used but allocation was not at random. Differences between intervention and control areas would be confounded by natural variability among areas, so the effect of the intervention could not be estimated.

Lack of sample size calculations

Only seven of the studies reported sample size calculations. Four were controlled trials in which the individual was the unit of allocation.[167,199,200,204] In the remaining three studies, calculations were also presented for the number of individuals to be sampled, from a single cluster,[198] from six hospitals,[176] and from four hospital departments.[183] The required number of clusters, or the number of individuals required, after adjusting for correlation of outcomes within clusters, were not reported. It has not been standard practice to report sample size calculations except for randomised trials, but the results of this survey suggest that methods to estimate the number of organisational clusters for inclusion in healthcare evaluation studies are generally not used, even though this may have a critical influence on the outcome of the evaluation. An example of an appropriate sample size calculation is given in chapter 5.

Individual level analysis without adjustment for correlation of responses within areas or organisations

A common feature of these studies was the reporting of individual level analyses without adjustment for correlation of responses within organisational clusters. This was not a surprising finding, as this was standard practice in the absence of suitable statistical software for this type of analysis.

Example: does a shorter length of hospital stay affect the outcome and costs of hysterectomy in southern England?[176]

This study was designed to see whether a shorter length of postoperative stay was associated with health outcomes after abdominal hysterectomy. Data were analysed for 363 women attending six hospitals. Conventional individual level multiple regression analyses were carried out, but as outcome variables might have been correlated within hospitals, it would now be advisable to allow for clustering of responses within organisational units by using regression methods for clustered data described in chapter 6. The application of regression methods for clustered data would generally require a larger number of hospitals than were included in this study.

Inappropriate methods used to allow for clustering of responses

An approach to allowing for clustering of responses that was common in the past was to include a fixed effect indicator variable for each cluster in multiple regression analyses. This was exemplified by eight studies in this review, for example.[163] The limitations of this approach are summarised in chapter 6.

Appropriate methods of individual level analysis of clustered data

When the number of clusters is small, it may be possible to report within-cluster analyses separately for each cluster. A study of the appropriateness of hospitalisation in a Spanish hospital reported some of the results separately for each of four hospital departments, before going on to report an overall individual level analysis.[177] Another appropriate method for individual level statistical analysis is to use the method of GEEs to adjust standard errors for the correlation structure of the data.[152] A study which examined the effects of patient volume and level of care at the hospital of birth on neonatal mortality reported a series of regression analyses in which the unit of observation was each birth. The method of GEEs was used to adjust standard errors for within hospital correlation.[152] The GEE approach is now being increasingly used in the analysis of healthcare evaluations.[151,212] However, the method is may prove unsatisfactory when the number of clusters is small.

Inclusion of cluster level covariates in individual level analyses

A further problem with individual level analysis is the difficulty of including cluster level covariates.

These are sometimes included as individual level characteristics as, for example, when the grade of surgeon carrying out a surgical operation is included at the individual level.[176] This approach would now be considered incorrect because the confidence intervals for the covariate will be erroneously small. It would be preferable to use a statistical technique which allowed the inclusion of covariates at both individual and cluster level.[123] In the present survey, an example of two-stage, two-level analysis was provided by a study of hospital- and patient-related characteristics determining length of hospital stay for hip and knee replacements.[179] Another study used GEE methods to adjust the standard errors of estimates of hospital level variables for within-hospital correlation of the outcome.[152]

Cluster level analysis without inclusion of individual level covariates

Cluster level analysis has the advantage that it will provide valid results in the presence of between-cluster variation, but analysis at the cluster level suffers from the disadvantage that individual level covariates cannot be included directly. However, standardisation techniques may be used to adjust for individual level characteristics such as age and sex.

Example of cluster level analysis without individual level covariates: influences of practice characteristics on prescribing in fundholding and non-fundholding general practices

Wilson and colleagues[206] studied practice level prescribing data for 384 practices in the former Mersey region of England. Regression analyses were carried out to determine whether general practice prescribing was associated with fundholding status. Analyses included a range of cluster level covariates, such as whether the practice was a training practice and whether the GP was practising single handed. Patient-specific characteristics such as diagnoses, and measures of comorbidity or disability, which might affect prescribing, could not be included directly.

Example of cluster level analysis after standardisation for individual level characteristics: challenges of monitoring use of secondary care at local level – a study based in London, UK[173]

Chenet and McKee analysed rates of hospital utilisation at electoral ward level in London. In order to allow for the varying age structure of ward populations, indirect standardisation was used to estimate an age-standardised hospital episode rate. The observed number of hospital episodes in each ward was expressed as a ratio compared with the number expected if age-specific hospital episode rates for the whole area were applied to the ward population.

Conclusions

The illustrations provided by this review suggest that theoretical developments in the evaluation of cluster-based interventions are not yet being widely applied in the practice of healthcare evaluation. Current practice could be improved by adopting simple measures such as including a sufficient number of clusters, including control clusters, estimating sample size requirements, and using methods of analysis which allow for the correlation structure of the data (see *Box 4*). The key first step is to recognise that clusters of individuals, rather than individuals themselves, are the units of intervention and evaluation.

The methods used in the evaluations reviewed here generally differ from those which were described in earlier sections of the review. At first sight it might appear that the focus of our review differed from the focus of this case study. Instead, we suggest that the findings of the case study reflect the extent to which existing practice in healthcare evaluation departs from what is justified in terms of recent methodological development.

Chapter 9

Database of intraclass correlation coefficients and variance components

Introduction

The need for rigorous evaluation of interventions directed at individual patients or healthy subjects, through the use of randomised trials, is well accepted. Public health interventions may be directed not at individuals but at geographical areas, communities or units of healthcare organisation. The need for evaluation of public health policies, programmes and interventions is being increasingly recognised, but the special difficulties encountered in implementing evaluations of public health interventions are not always adequately appreciated.[9] These types of evaluation present distinct problems because the unit of allocation is not an individual subject but a cluster of individuals such as a general practice population, a hospital, or a geographical or administrative area.

One of the key differences between evaluations at cluster level rather than individual level is the dependence of individual observations within clusters.[109] Individuals sharing the same geographical area or organisational unit tend to be more similar to each other than to individuals in other areas. If cluster level outcomes are being considered, this is not a problem. But cluster level interventions are often aimed at modifying individual level outcomes. For example, a community-wide health promotion programme may be implemented with the objective of reducing risk factors for cardiovascular disease in individuals. When individual level analyses are to be carried out, between-cluster variation contributes an additional source of variation which must be allowed for, in addition to between-subject, within-cluster variation. When between-cluster variation is present, the number of individuals needed for a cluster-based study is larger than for a study of the same power in which individuals are allocated.[109]

In order to estimate the required sample size, the design effect must be incorporated into standard sample size formulae.[108,109,111] The design effect is a function of the average cluster size and the intraclass correlation coefficient:[84]

$$\text{design effect} = 1 + (n - 1)\rho$$

where n is the average cluster size (or average number of individuals sampled per cluster), and ρ is the intraclass correlation coefficient of the outcome. ρ quantifies the extent of between-cluster variation. It represents the proportion of the true total variation in the outcome that can be attributed to differences between the clusters:

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

where σ_b^2 is the between-cluster variance component, and σ_w^2 is the within-cluster variance component.

One of the problems investigators face in designing cluster-based evaluations is that estimates of ρ need to be obtained from previous studies. Nearly every author commenting on this subject has called for the publication of intraclass correlation coefficients which could be used to aid the design of future area- or organisation-based studies. A few reports have recently appeared providing data for US populations,[103,213] but equivalent data appear not to have been reported in Britain. As part of this methodological systematic review, we estimated intraclass correlations from data obtained from a number of different sources. In this section of the report we present the methods used and the results obtained for different geographical and organisational levels of clustering.

Methods

Data

Data were obtained from a number of sources, which can be classified under the following headings.

Health Survey for England 1994

The Health Survey for England is a health and lifestyle survey carried out annually in England. Data collection procedures include interview, anthropometric measurement and blood sampling.

Data from the Health Survey for England 1994 (Crown copyright 1994, used by permission of the Office for National Statistics) were obtained from the Data Archive, University of Essex. The design of the Health Survey for England was reported by Colhoun and Prescott-Clarke.[214] A multistage sampling framework was used, with postcode sectors as the primary sampling unit, and households sampled from each of the postcode sectors. Sampling of postcode sectors was stratified by regional health authority as well as by four socio-demographic variables: the proportion aged 65 years or more; the proportion of households without a car; the proportion of economically active males unemployed; and the proportion of non-white adults. Within households, all adults aged 16 years or more were eligible for inclusion in the survey.

The data collection procedures of the Health Survey for England and definitions of key variables are given by Colhoun and Prescott Clarke.[214] For these analyses, we selected a range of variables which might be considered as potential outcome variables for future community intervention studies. These included analytes obtained from a blood sample: the total serum cholesterol concentration (mmol/l); glycated haemoglobin (%); plasma fibrinogen (g/l); serum ferritin (μg/l); and haemoglobin concentration (g/dl). We also included the body mass index (BMI; as weight (kg)/height (m^2)) and whether subjects were overweight (BMI > 25 kg/m^2) or obese (BMI > 30 kg/m^2). Other data were obtained by interview. Where appropriate, the categories of qualitative variables were reduced to those shown in the tables. Each variable should be understood to contrast with the condition not present; for example, 'eats fruit at least once a day or not'. The variables shown in the tables are mostly self-explanatory, but more complete definitions are given by Colhoun and Prescott-Clarke.[214] The item 'drinks more than recommended limit' was derived as the proportion drinking more than 21 units per week for men or 14 units per week for women.

British Regional Heart Study

The design of the British Regional Heart Study was described by Shaper and co-workers[215] In brief, 24 towns were selected in England, Wales and Scotland in order to represent areas with differing mortality from cardiovascular disease. In each town, a general practice was selected and men aged 40–59 years, who were registered with the practice, were included in the study. Data were obtained from town profiles provided by the Department of Clinical Epidemiology and General Practice at the Royal Free Hospital School of Medicine. These were based on data obtained from 7735 men who were examined in 1978–1980.

It should be emphasised that the towns included in the study were not randomly sampled but were selected because they had different levels of coronary heart disease mortality.

Public Health Common Data Set

The Public Health Common Data Set for 1996 was obtained from the UK Department of Health. This data set includes mortality and morbidity data at district health authority level for England in 1994.

Health service indicators 1996

Health service indicators for the UK National Health Service in 1996 were obtained from the National Health Service Executive. This data set includes health service activity data aggregated to district health authority, family health service authority and acute unit (hospital) level.

Thames Cancer Registry

Cancer incidence data were obtained from the Thames Cancer Registry for the 27 District Health Authorities in the former South Thames Regions.

Other

Other data included in the database were abstracted from published papers. Data were also analysed for the Royal College of Physicians Wound Care Audit.[216] We also made contact with a number of other investigators and audit groups. However, because of considerations of confidentiality, accessing other sources of unpublished data proved difficult, and the database does not currently contain a wide range of data for general practice level outcomes.

Analysis

We estimated components of variance for continuous variables using analysis of variance. For continuous variables, approximate confidence intervals for ρ may be estimated when the design is unbalanced, as described by Donner and Wells,[98] but these methods have so far only been evaluated for smaller clusters such as families. We have therefore not included confidence intervals for ρ. For binary variables, individual level data were analysed by use of analysis of variance. When binary data were available aggregated to cluster level, kappa was estimated.[91] In the case of binary

variables, it is appropriate to include a dichotomous outcome as the dependent variable in the analysis of variance, but because the distributional assumptions are not met, it is not appropriate to attempt interval estimation. Data presented include the average cluster size, the number of clusters, the overall mean for continuous variables or overall prevalence for binary variables, the within-cluster and between-cluster components of variance, ρ and the design effect. Where the between-cluster component of variance was negative, ρ was truncated at zero. Data are presented in the same format as earlier papers, to five decimal places.[103,213]

For analysis of data from the Health Survey for England, we considered that individual values were clustered within households, which in turn were clustered within postcode sectors. Postcode sectors were considered to be clustered within district and regional health authorities. We have not allowed for the stratified sampling of postcode sectors within regional health authorities; postcode sector level data were not available for analysis. Analyses were performed using the procedure NESTED in SAS.[100] This procedure performs a random effects analysis of variance which is appropriate for a multistage nested sampling design. The NESTED procedure is computationally more efficient than alternative procedures in SAS such as GLM and MIXED, resulting in greatly reduced analysis times.

We used a four-level completely nested model to estimate components of variance for each level of clustering:

$$y_{ijklr} = \mu + \alpha_i + \beta_{j(i)} + \gamma_{k(ij)} + \delta_{l(ijk)} + \epsilon_{r(ijkl)}$$

where y_{ijklr} is the value of the dependent variable observed for the rth subject, in the lth household, kth postcode sector, jth district health authority and ith regional health authority, μ is the overall mean of the sampled population, and α_i, $\beta_{j(i)}$, $\gamma_{k(ij)}$, $\delta_{l(ijk)}$ and $\epsilon_{r(ijkl)}$ are uncorrelated random effects with zero means and respective variances σ^2_{rha}, σ^2_{dha}, σ^2_{pcs}, σ^2_{hh} and σ^2_e (subscripts: rha, regional health authority; dha, district health authority; pcs, postcode sector; hh, households). We used variance components from the four level model to estimate the following intraclass correlations:

$$\rho_{rha} = \sigma^2_{rha} / (\sigma^2_{rha} + \sigma^2_{dha} + \sigma^2_{pcs} + \sigma^2_{hh} + \sigma^2_e)$$
$$\rho_{dha} = \sigma^2_{dha} / (\sigma^2_{dha} + \sigma^2_{pcs} + \sigma^2_{hh} + \sigma^2_e)$$
$$\rho_{pcs} = \sigma^2_{pcs} / (\sigma^2_{pcs} + \sigma^2_{hh} + \sigma^2_e)$$
$$\rho_{hh} = \sigma^2_{hh} / (\sigma^2_{hh} + \sigma^2_e)$$

For each equation, the numerator represents the between-cluster component of variance, and the difference between the numerator and denominator represents the within-cluster component of variance. In using this model we have assumed that if randomisation were to be carried out at a lower level, then variation at higher levels would be accounted for by use of stratification. For example, we have assumed that randomisation by district health authority would be stratified according to regional health authority. If regional health authority level variation were to be subsumed into district health authority level variation, then the district health authority component of variance would be correspondingly larger. The same models were used for continuous and binary variables. We emphasise that the design effects presented here differ from those presented in the report of the Health Survey for England 1994[214] (see Tables 13.16–13.33 therein). In our report we present the design effect to be expected if clusters of the same size were allocated in a cluster randomised study. The design effects included in the report of the Health Survey for England represent the overall design effect of the survey.

Results

Results are presented for cardiovascular diseases and lifestyle risk factors (*Tables 12–27*), cancer (*Table 28*), respiratory disease (*Tables 29–31*), health service activity (*Tables 32–41*) and other (*Tables 42–43*).

The most obvious feature of the results is the inverse relationship between cluster size and intraclass correlation coefficient, which is most apparent on inspection of the data from the Health Survey for England. At regional and district health authority level, ρ was generally below 0.01. For postcode sector level, ρ was generally less than 0.05, but at household level, ρ was mostly in the range 0.0–0.3. However, for larger cluster sizes, even small values of ρ might be associated with substantial design effects that could not be ignored in designing studies.

Table 25 shows ρ values separately for men and women, for continuous variables at postcode sector level. It can be seen that ρ showed some gender differences which differed among the variables. However, the result for both sexes combined was generally of approximately the same magnitude as for the sexes separately.

Discussion

In this section of the report we have presented components of variance and intraclass correlation coefficients for a range of outcomes which may be relevant in the design of area- or organisation-based studies. Only one data set was obtained from an experimental study (*Tables 30* and *31*). In general, experimental data will usually be obtained from a small number of clusters, giving imprecise estimates for intraclass correlations. The present analyses of observational data generally included sufficiently large numbers of clusters to allow intraclass correlations to be estimated with reasonable precision. For example, data were obtained from more than 7000 households and more than 700 postcode sectors.

Data from the Health Survey for England will be particularly relevant to the design of community-based studies. The types of unit of allocation which might be used in a community intervention study may vary, but the levels for which we have presented data may be broadly generalised. The household will generally correspond fairly closely to a family. In England there are 7223 postcode sectors with average populations of about 6500, these correspond fairly closely in size to electoral wards which have average populations of about 5200. District health authorities are administrative areas responsible for the administration of health-care services, and at the time of the 1994 survey had average populations of just over 250,000. District health authorities are commonly based on small towns, counties or sections of cities.

Estimating components of variance from observational data has the advantage that large numbers of clusters may be included, leading to more precise estimates of intraclass correlations. Furthermore, data from national sources may be considered more generalisable than data obtained through intervention in a single locality. Nevertheless, potential users of the present data will need to consider the extent to which the findings may be generalised to different settings and sampling designs. There may be a danger in extrapolating the intraclass correlation coefficient from one study to another because one or other of the components of variance (either the between-cluster or within-cluster variance) may vary from one population to another, or may at least be dependent on the sampling strategy. For this reason we have presented components of variance as well as intraclass correlations. We recommend that a sensitivity analysis be included when these data are used to aid the estimation of sample size requirements.

A second issue which deserves consideration is the extent to which intraclass correlation coefficients will be affected by varying age or sex distributions in the sampled populations. In order to simplify the data presentation for this report, we have presented data for both sexes and for all age groups combined. These estimates are likely to have the widest application, but it must be acknowledged that sampling according to age or sex might be expected to influence estimates of intraclass correlation. Nevertheless, our basic conclusion that between-cluster variation is usually appreciable, and must be considered, remains unaltered. Furthermore, intraclass correlation coefficients tend to be larger for lower levels of clustering, but for larger cluster sizes, design effects may be substantial even when the intraclass correlation coefficient is extremely small.

TABLE 12 *Cardiovascular and lifestyle 1. National health strategy target indicators*

Variable	Source of data	Setting	Unit type	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall pro-portion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Coronary heart disease mortality	PHCDS	1995	Men aged < 65 years	DHA	198,419	105	128.7	0.000648	1.8×10^{-8}	0.000647	0.0000283	6.62
Coronary heart disease mortality	PHCDS	1995	Women aged < 65 years	DHA	193,141	105	34.53	0.000179	3.0×10^{-9}	0.000179	0.0000145	3.80
Coronary heart disease mortality	PHCDS	1995	All aged < 65 years	DHA	391,561	105	163.2	0.000416	9.0×10^{-9}	0.000416	0.0000210	9.22
Coronary heart disease mortality	PHCDS	1995	Men aged 65–74 years	DHA	18,524	105	201.5	0.0109	3.0×10^{-6}	0.0107	0.000277	6.13
Coronary heart disease mortality	PHCDS	1995	Women aged 65–74 years	DHA	22,084	105	104.5	0.00472	1.0×10^{-6}	0.00470	0.000212	5.68
Coronary heart disease mortality	PHCDS	1995	All aged 65–74 years	DHA	40,608	105	306.0	0.00752	1.8×10^{-6}	0.00747	0.000235	10.5
Stroke mortality	PHCDS	1995	Men aged < 65 years	DHA	198,419	105	22.7	0.000114	$< 10^{-9}$	0.000114	0.00000419	1.83
Stroke mortality	PHCDS	1995	Women aged < 65 years	DHA	193,141	105	17.8	0.0000921	$< 10^{-9}$	0.0000921	0.00000269	1.52
Stroke mortality	PHCDS	1995	All aged < 65 years	DHA	391,561	105	40.5	0.000103	$< 10^{-9}$	0.000103	0.00000439	2.72
Stroke mortality	PHCDS	1995	Men aged 65–74 years	DHA	18,524	105	45.4	0.00245	1.9×10^{-7}	0.00244	0.0000764	2.42
Stroke mortality	PHCDS	1995	Women aged 65–74 years	DHA	22,084	105	42.3	0.00191	9.4×10^{-8}	0.00191	0.0000490	2.08
Stroke mortality	PHCDS	1995	All aged 65–74 years	DHA	40,608	105	87.7	0.00216	1.2×10^{-7}	0.00215	0.0000576	3.34

PHCDS, Public Health Data Set; DHA, district health authority

TABLE 13 *Cardiovascular and lifestyle 2. Data from the Health Survey for England (HSE) 1994.*[214] *Continuous variables at regional health authority (RHA) level*

Variable	Source	Setting	Cluster type	Average cluster size	Number of clusters	Overall mean	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Serum total cholesterol (mmol/l)	HSE	England	RHA	790	14	5.88	0.00352	1.63865	0.00214	2.69
Glycated haemoglobin (%)	HSE	England	RHA	774	14	6.44	0.00000	1.14775	0.00000	1.00
Plasma fibrinogen (g/l)	HSE	England	RHA	693	14	3.10	0.00082	0.68725	0.00119	1.82
Serum ferritin (µg/l)	HSE	England	RHA	778	14	76.22	0.26739	6693.10	0.00004	1.03
Haemoglobin level (g/l)	HSE	England	RHA	764	14	13.96	0.00127	2.16567	0.00059	1.45
Systolic blood pressure (mmHg)	HSE	England	RHA	893	14	135.79	1.10876	404.73	0.00273	3.44
Diastolic blood pressure (mmHg)	HSE	England	RHA	893	14	74.43	0.20512	161.25	0.00127	2.13
BMI (kg/m^2)	HSE	England	RHA	1044	14	25.89	0.03560	20.435	0.00174	2.81
Waist circumference (cm)	HSE	England	RHA	946	14	87.8	0.40276	163.980	0.00245	3.32
Hip circumference (cm)	HSE	England	RHA	948	14	103.97	0.26481	84.779	0.00311	3.95

TABLE 14 *Cardiovascular and lifestyle 3. Data from the HSE 1994.[214] Categorical variables at RHA level*

Variable	Source	Setting	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall proportion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Drinks more than recommended limit	HSE	England	RHA	1123	14	238	0.209	0.00063	0.16518	0.00378	5.24
Ever smoked cigarettes	HSE	England	RHA	1122	14	609	0.541	0.00000	0.24851	0.00000	1.00
Current cigarette smoker	HSE	England	RHA	1122	14	345	0.306	0.00000	0.21263	0.00000	1.00
Current smoker or passive smoke exposure	HSE	England	RHA	1122	14	473	0.419	0.00000	0.24388	0.00000	1.00
Had GP consultation in last 14 days	HSE	England	RHA	1121	14	192	0.171	0.00007	0.14140	0.00051	1.57
On contraceptive pill (menstruating women only)	HSE	England	RHA	331	14	86	0.259	0.00000	0.19200	0.00000	1.00
Moderately active at home/ in garden	HSE	England	RHA	1124	14	766	0.679	0.00043	0.21775	0.00197	3.21
Active in sport – moderate/vigorous	HSE	England	RHA	1124	14	482	0.427	0.00063	0.24417	0.00255	3.87
Active in general – moderate/vigorous	HSE	England	RHA	1124	14	939	0.832	0.00015	0.13978	0.00110	2.24
Physically inactive	HSE	England	RHA	1124	14	642	0.569	0.00010	0.24540	0.00042	1.47
Active at work – moderate/vigorous	HSE	England	RHA	1118	14	165	0.147	0.00003	0.12693	0.00024	1.27

TABLE 15 *Cardiovascular and lifestyle 4. Data from the HSE 1994.[214] Categorical variables at RHA level*

Variable	Source	Setting	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall proportion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Overweight (BMI > 25 kg/m^2)	HSE	England	RHA	1044	14	557	0.531	0.00026	0.24884	0.00105	2.10
Obese (BMI > 30 kg/m^2)	HSE	England	RHA	1044	14	164	0.157	0.00017	0.13231	0.00132	2.38
Eats fruit at least once a day	HSE	England	RHA	1121	14	560	0.497	0.00165	0.24850	0.00658	8.37
Eats vegetables at least once a day	HSE	England	RHA	1117	14	760	0.677	0.00392	0.21506	0.01792	21.00
Adds salt to food when cooking	HSE	England	RHA	1107	14	759	0.683	0.00033	0.21623	0.00152	2.68
Adds salt to meal	HSE	England	RHA	1123	14	613	0.544	0.00151	0.24673	0.00609	7.83
Doctor-diagnosed diabetes	HSE	England	RHA	1123	14	27	0.024	0.00001	0.02338	0.00043	1.48
Currently has high blood pressure	HSE	England	RHA	1088	14	107	0.098	0.00000	0.08357	0.00000	1.00
Doctor-diagnosed angina	HSE	England	RHA	1124	14	43	0.038	0.00000	0.03715	0.00000	1.00
Doctor-diagnosed heart attack	HSE	England	RHA	1124	14	30	0.027	0.00000	0.02605	0.00000	1.00
Doctor-diagnosed stroke	HSE	England	RHA	1124	14	19	0.017	0.00000	0.01647	0.00000	1.00
Doctor-diagnosed ischaemic heart disease (angina/ heart attack)	HSE	England	RHA	1124	14	56	0.050	0.00000	0.04743	0.00000	1.00

TABLE 16 *Cardiovascular and lifestyle 5. Data from the HSE 1994.[214] Continuous variables at district health authority (DHA) level*

Variable	Source	Setting	Cluster type	Average cluster size	Number of clusters	Overall mean	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Serum total cholesterol (mmol/l)	HSE	England	DHA	61	177	5.88	0.00406	1.63464	0.00244	1.15
Glycated haemoglobin (%)	HSE	England	DHA	60	177	6.44	0.00634	1.14141	0.00552	1.33
Plasma fibrinogen (g/l)	HSE	England	DHA	53	177	3.10	0.00000	0.68725	0.00000	1.00
Serum ferritin (µg/l)	HSE	England	DHA	60	177	76.22	0.00000	6693.1	0.00000	1.00
Haemoglobin (g/dl)	HSE	England	DHA	59	177	13.96	0.00499	2.16069	0.00271	1.16
Systolic blood pressure (mmHg)	HSE	England	DHA	69	177	135.79	2.94351	401.788	0.00727	1.49
Diastolic blood pressure (mmHg)	HSE	England	DHA	69	177	74.43	1.05226	160.196	0.00653	1.44
BMI (kg/m^2)	HSE	England	DHA	81	177	28.89	0.02876	20.407	0.00141	1.11
Waist circumference (cm)	HSE	England	DHA	73	177	87.77	0.11034	163.869	0.00067	1.05
Hip circumference (cm)	HSE	England	DHA	73	177	103.97	0.20623	84.572	0.00243	1.17

TABLE 17 *Cardiovascular and lifestyle 6. Data from the HSE 1994.[214] Categorical variables at DHA level*

Variable	Source	Setting	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall pro-portion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Drinks more than recommended limit	HSE	England	DHA	87	177	18.71	0.209	0.00079	0.16440	0.00476	1.41
Ever smoked cigarettes	HSE	England	DHA	87	177	48.17	0.541	0.00098	0.24753	0.00395	1.34
Current cigarette smoker	HSE	England	DHA	87	177	27.28	0.306	0.00151	0.21112	0.00711	1.61
Current smoker or passive exposure	HSE	England	DHA	87	177	37.42	0.419	0.00247	0.24141	0.01012	1.87
Had GP consultation in last 14 days	HSE	England	DHA	87	177	15.19	0.171	0.00037	0.14104	0.00258	1.22
On contraceptive pill (menstruating women only)	HSE	England	DHA	26	177	6.80	0.259	0.00225	0.19521	0.01139	1.28
Moderately active at home/ in garden	HSE	England	DHA	87	177	60.60	0.679	0.00000	0.21775	0.00000	1.00
Active in sport – moderate/vigorous	HSE	England	DHA	87	177	38.15	0.427	0.00190	0.24228	0.00777	1.67
Active in general – moderate/vigorous	HSE	England	DHA	87	177	74.29	0.832	0.00053	0.13925	0.00379	1.33
Physically inactive	HSE	England	DHA	87	177	50.79	0.569	0.00000	0.24540	0.00000	1.00
Active at work – moderate/vigorous	HSE	England	DHA	86	177	13.056	0.147	0.00039	0.12653	0.00307	1.26

TABLE 18 *Cardiovascular and lifestyle 7. Data from the HSE 1994.[214] Categorical variables at DHA level*

Variable	Source	Setting	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall pro-portion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Overweight (BMI > 25 kg/m^2)	HSE	England	DHA	81	177	44.02	0.531	0.00054	0.24830	0.00216	1.17
Obese (BMI > 30 kg/m^2)	HSE	England	DHA	81	177	12.99	0.157	0.00000	0.13231	0.00000	1.00
Eats fruit at least once a day	HSE	England	DHA	87	177	44.28	0.497	0.00007	0.24842	0.00029	1.02
Eats vegetables at least once a day	HSE	England	DHA	86	177	60.10	0.677	0.00081	0.21425	0.00373	1.32
Adds salt to food when cooking	HSE	England	DHA	85	177	60.06	0.683	0.00177	0.21446	0.00818	1.69
Adds salt to meal	HSE	England	DHA	87	177	48.47	0.544	0.00061	0.24611	0.00247	1.21
Doctor-diagnosed diabetes	HSE	England	DHA	87	177	2.14	0.024	0.00000	0.02338	0.00000	1.00
Currently has high blood pressure	HSE	England	DHA	84	177	8.47	0.098	0.00007	0.08350	0.00084	1.07
Doctor-diagnosed angina	HSE	England	DHA	87	177	3.42	0.038	0.00003	0.03712	0.00081	1.07
Doctor-diagnosed heart attack	HSE	England	DHA	87	177	2.37	0.027	0.00004	0.02601	0.00154	1.13
Doctor-diagnosed stroke	HSE	England	DHA	87	177	1.49	0.017	0.00000	0.01647	0.00000	1.00
Doctor-diagnosed ischaemic heart disease (angina/heart attack)	HSE	England	DHA	87	177	4.42	0.050	0.00005	0.04738	0.00105	1.09

TABLE 19 *Cardiovascular and lifestyle 8. Data from the HSE 1994.[214] Continuous variables at postcode sector level*

Variable	Source	Setting	Cluster type	Average cluster size	Number of clusters	Overall mean	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Total cholesterol (mmol/l)	HSE	England	Postcode sector	17	711	5.88	0.03788	1.59677	0.02317	1.37
Glycated haemoglobin (g/dl)	HSE	England	Postcode sector	15	711	6.44	0.02488	1.11653	0.02180	1.31
Fibrinogen (g/l)	HSE	England	Postcode sector	13	711	3.10	0.03737	0.64989	0.05437	1.65
Ferritin (µg/l)	HSE	England	Postcode sector	15	711	76.22	93.2586	6599.84	0.01393	1.20
Haemoglobin (g/dl)	HSE	England	Postcode sector	15	711	13.96	0.05883	2.10186	0.02723	1.38
Systolic blood pressure (mmHg)	HSE	England	Postcode sector	17	711	135.79	7.69267	394.095	0.01915	1.31
Diastolic blood pressure (mmHg)	HSE	England	Postcode sector	17	711	74.43	3.30140	156.895	0.02061	1.33
BMI (kg/m^2)	HSE	England	Postcode sector	20	712	25.887	0.09116	20.3156	0.00447	1.08
Waist circumference (cm)	HSE	England	Postcode sector	18	711	87.769	3.26321	160.606	0.01991	1.34
Hip circumference (cm)	HSE	England	Postcode sector	19	711	103.967	1.56795	83.005	0.01854	1.33

TABLE 20 *Cardiovascular and lifestyle 9. Data from the HSE 1994.[214] Categorical variables at postcode sector level*

Variable	Source	Setting	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall pro-portion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Drinks more than recommended limit	HSE	England	Postcode sector	22	712	4.65	0.209	0.00138	0.16302	0.00841	1.18
Ever smoked cigarettes	HSE	England	Postcode sector	22	712	11.97	0.541	0.00112	0.24640	0.00453	1.10
Current cigarette smoker	HSE	England	Postcode sector	22	712	6.78	0.306	0.00089	0.21023	0.00421	1.09
Current smoker or passive exposure	HSE	England	Postcode sector	22	712	9.30	0.419	0.00150	0.23991	0.00620	1.13
Had GP consultation in last 14 days	HSE	England	Postcode sector	22	712	3.77	0.171	0.00041	0.14063	0.00288	1.06
On contraceptive pill (menstruating women only)	HSE	England	Postcode sector	6	709	1.70	0.259	0.00142	0.19378	0.00727	1.04
Moderately active at home/ in garden	HSE	England	Postcode sector	22	712	15.06	0.679	0.00372	0.21403	0.01710	1.36
Active in sport – moderate/vigorous	HSE	England	Postcode sector	22	712	9.48	0.427	0.00323	0.23904	0.01334	1.28
Active in general – moderate/vigorous	HSE	England	Postcode sector	22	712	18.47	0.832	0.00098	0.13827	0.00702	1.15
Physically inactive	HSE	England	Postcode sector	22	712	12.63	0.569	0.00265	0.24275	0.01079	1.23
Active at work – moderate/vigorous	HSE	England	Postcode sector	22	712	3.25	0.147	0.00156	0.12497	0.01233	1.26

TABLE 21 *Cardiovascular and lifestyle 10. Data from the HSE 1994.[214] Categorical variables at postcode sector level*

Variable	Source	Setting	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall pro-portion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Overweight (BMI > 25 kg/m^2)	HSE	England	Postcode sector	20	712	10.94	0.531	0.00128	0.24702	0.00516	1.10
Obese (BMI > 30 kg/m^2)	HSE	England	Postcode sector	20	712	3.23	0.157	0.00021	0.13210	0.00159	1.03
Eats fruit at least once a day	HSE	England	Postcode sector	22	712	11.01	0.497	0.00272	0.24570	0.01095	1.23
Eats vegetables at least once a day	HSE	England	Postcode sector	22	712	14.94	0.677	0.00577	0.20849	0.02692	1.57
Adds salt to food when cooking	HSE	England	Postcode sector	22	712	14.93	0.683	0.00195	0.21251	0.00908	1.19
Adds salt to meal	HSE	England	Postcode sector	22	712	12.05	0.544	0.00129	0.24482	0.00524	1.11
Doctor-diagnosed diabetes in subject	HSE	England	Postcode sector	22	712	0.531	0.0239	0.00002	0.02336	0.00085	1.02
Patient currently has high blood pressure	HSE	England	Postcode sector	21	712	2.11	0.0980	0.00000	0.08350	0.00000	1.00
Doctor-diagnosed angina	HSE	England	Postcode sector	22	712	0.851	0.0383	0.00000	0.03712	0.00000	1.00
Doctor-diagnosed heart attack	HSE	England	Postcode sector	22	712	0.590	0.0266	0.00000	0.02601	0.00000	1.00
Doctor-diagnosed stroke	HSE	England	Postcode sector	22	712	0.371	0.0167	0.00000	0.01647	0.00000	1.00
Doctor-diagnosed ischaemic heart disease (angina/heart attack)	HSE	England	Postcode sector	22	712	1.10	0.0495	0.00000	0.04738	0.00000	1.00

TABLE 22 *Cardiovascular and lifestyle 11. Data from the HSE 1994.[214] Continuous variables at household level*

Variable	Source	Setting	Cluster type	Average cluster size	Number of clusters	Overall mean	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Total cholesterol (mmol/l)	HSE	England	Household	1.57	6948	5.88	0.27762	1.31914	0.17387	1.10
Glycated haemoglobin (%)	HSE	England	Household	1.56	6874	6.44	0.24273	0.87380	0.21740	1.12
Fibrinogen (g/l)	HSE	England	Household	1.49	6413	3.10	0.20389	0.44599	0.31374	1.15
Ferritin (µg/l)	HSE	England	Household	1.56	6891	76.22	332.678	6267.16	0.05041	1.03
Haemoglobin (g/dl)	HSE	England	Household	1.55	6826	13.96	0.00000	2.10186	0.00000	1.00
Systolic blood pressure (mmHg)	HSE	England	Household	1.65	7487	135.79	143.577	250.518	0.36432	1.24
Diastolic blood pressure (mmHg)	HSE	England	Household	1.65	7487	74.43	30.6935	126.201	0.19563	1.13
BMI (kg/m^2)	HSE	England	Household	1.69	8559	25.887	3.95798	16.3576	0.19482	1.13
Waist circumference (cm)	HSE	England	Household	1.69	7764	87.769	3.54565	157.060	0.02208	1.02
Hip circumference (cm)	HSE	England	Household	1.69	7766	103.967	16.4556	66.5489	0.19825	1.14

TABLE 23 *Cardiovascular and lifestyle 12. Data from the HSE 1994.[214] Categorical variables at household level*

Variable	Source	Setting	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall pro-portion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Drinks more than recommended limit	HSE	England	Household	1.72	9063	0.365	0.209	0.03238	0.13063	0.19864	1.14
Ever smoked cigarettes	HSE	England	Household	1.72	9060	0.941	0.541	0.04246	0.20394	0.17232	1.12
Current cigarette smoke	HSE	England	Household	1.72	9059	0.533	0.306	0.06064	0.14959	0.28845	1.21
Current smoker or passive exposure	HSE	England	Household	1.72	9059	0.731	0.419	0.20912	0.03080	0.87163	1.63
Had GP consultation in last 14 days	HSE	England	Household	1.72	9049	0.297	0.171	0.01235	0.12828	0.08781	1.06
On contraceptive pill (menstruating women only)	HSE	England	Household	1.09	4190	0.287	0.259	0.00000	0.19379	0.00000	1.00
Moderately active at home/ in garden	HSE	England	Household	1.72	9067	1.18	0.679	0.02138	0.19265	0.09991	1.07
Active in sport – moderate/vigorous	HSE	England	Household	1.72	9067	0.745	0.427	0.06138	0.17767	0.25676	1.18
Active in general – moderate/vigorous	HSE	England	Household	1.72	9067	1.450	0.832	0.03467	0.10360	0.25071	1.18
Physically inactive	HSE	England	Household	1.72	9067	0.991	0.569	0.03539	0.20737	0.14577	1.10
Active at work – moderate/vigorous	HSE	England	Household	1.71	9060	0.255	0.147	0.00000	0.12497	0.00000	1.00

TABLE 24 *Cardiovascular and lifestyle 13. Data from the HSE 1994.[214] Categorical variables at household level*

Variable	Source	Setting	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall proportion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Overweight (BMI > 25 kg/m^2)	HSE	England	Household	1.69	8559	0.910	0.531	0.02934	0.21768	0.11877	1.08
Obese (BMI > 30 kg/m^2)	HSE	England	Household	1.69	8559	0.269	0.157	0.01433	0.11777	0.10846	1.07
Eats fruit at least once a day	HSE	England	Household	1.72	9053	0.866	0.497	0.08068	0.16503	0.32835	1.24
Eats vegetables at least once a day	HSE	England	Household	1.71	9036	1.177	0.677	0.11130	0.09718	0.53386	1.38
Adds salt to food when cooking	HSE	England	Household	1.71	8984	1.183	0.683	0.16931	0.04321	0.79669	1.57
Adds salt to meal	HSE	England	Household	1.72	9058	0.947	0.544	0.05940	0.18542	0.24263	1.17
Doctor-diagnosed diabetes	HSE	England	Household	1.72	9064	0.0417	0.0239	0.00177	0.02159	0.07608	1.05
Currently has high blood pressure	HSE	England	Household	1.69	8923	0.168	0.0980	0.02042	0.06831	0.23011	1.16
Doctor-diagnosed angina	HSE	England	Household	1.72	9067	0.0668	0.0383	0.00907	0.02805	0.24440	1.18
Doctor-diagnosed heart attack	HSE	England	Household	1.72	9067	0.0463	0.0266	0.00365	0.02236	0.14031	1.10
Doctor-diagnosed stroke	HSE	England	Household	1.72	9067	0.0291	0.0167	0.00339	0.01308	0.20573	1.15
Doctor-diagnosed ischaemic heart disease (angina/heart attack)	HSE	England	Household	1.72	9067	0.0864	0.0495	0.01105	0.03633	0.23324	1.17

TABLE 25 *Cardiovascular and lifestyle 14. Data from the HSE 1994.[214] Continuous variables at postcode sector level by gender*

Variable	Intraclass correlation coefficient		
	All	Men	Women
Serum total cholesterol (mmol/l)	0.02317	0.03776	0.02232
Glycated haemoglobin (%)	0.02180	0.03153	0.01221
Plasma fibrinogen (g/l)	0.05437	0.06502	0.04483
Serum ferritin (µg/l)	0.01393	0.00639	0.00000
Haemoglobin (g/dl)	0.03205	0.04349	0.04876
Systolic blood pressure (mmHg)	0.01915	0.01994	0.02708
Diastolic blood pressure (mmHg)	0.02071	0.01854	0.02829
BMI (kg/m^2)	0.00447	0.01525	0.01187
Waist circumference (cm)	0.01991	0.03044	0.02925
Hip circumference (cm)	0.01854	0.03575	0.01736

TABLE 26 *Cardiovascular and lifestyle 15. Data from the British Regional Heart Study (BRHS)*[a 215]

Variable	Source	Setting	Unit type	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall proportion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Cigarette smoker	BRHS	England, Wales and Scotland	Men aged 40–59 years	Town	322.2	24	132.71	0.41	0.0063	0.24	0.026	9.35
Drinks more than 6 units on weekend	BRHS	England, Wales and Scotland	Men aged 40–59 years	Town	322.2	24	45.63	0.142	0.00439	0.117	0.0361	12.59
Drinks more than 6 units daily	BRHS	England, Wales and Scotland	Men aged 40–59 years	Town	322.2	24	34.67	0.108	0.00296	0.0932	0.0308	10.89
High blood pressure (SBP > 160 mmHg, DBP > 90 mmHg)	BRHS	England, Wales and Scotland	Men aged 40–59 years	Town	322.2	24	45.88	0.142	0.00208	0.120	0.0171	6.49
On antihypertensives	BRHS	England, Wales and Scotland	Men aged 40–59 years	Town	322.2	24	15.63	0.0485	0.000281	0.0459	0.00610	2.96
On anticoagulants	BRHS	England, Wales and Scotland	Men aged 40–59 years	Town	322.2	24	1.71	0.00530	0.00000756	0.00527	0.00143	1.46
On lipid-lowering drugs	BRHS	England, Wales and Scotland	Men aged 40–59 years	Town	322.2	24	1.54	0.00478	0.00000620	0.00476	0.00130	1.42

SBP, systolic blood pressure; DBP, diastolic blood pressure

[a] *Towns in the BRHS were not randomly sampled*

TABLE 27 *Cardiovascular and lifestyle 16. Data from the BRHS*[a] [215]

Variable	Source	Setting	Unit type	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall pro-portion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Angina (doctor diagnosed)	BRHS	England, Wales and Scotland	Men aged 40–59 years	Town	322.2	24	10.46	0.0324	0.0000955	0.0313	0.00304	1.98
Heart attack (doctor diagnosed)	BRHS	England, Wales and Scotland	Men aged 40–59 years	Town	322.2	24	12.13	0.0376	0.000149	0.0361	0.00411	2.32
High blood pressure (doctor diagnosed)	BRHS	England, Wales and Scotland	Men aged 40–59 years	Town	322.2	24	41.25	0.128	0.000558	0.111	0.00500	2.61
Stroke (doctor diagnosed)	BRHS	England, Wales and Scotland	Men aged 40–59 years	Town	322.2	24	2.29	0.00711	0.00000497	0.00706	0.000704	1.23
Diabetes (doctor diagnosed)	BRHS	England, Wales and Scotland	Men aged 40–59 years	Town	322.2	24	4.92	0.0153	0.0	0.0150	0.0	1.0
On oral antidiabetics	BRHS	England, Wales and Scotland	Men aged 40–59 years	Town	322.2	24	1.83	0.00569	0.0000111	0.00565	0.00195	1.63
On insulin injections	BRHS	England, Wales and Scotland	Men aged 40–59 years	Town	322.2	24	1.5	0.00465	0.0	0.00464	0.0	1.0
5 year incidence of fatal ischaemic heart disease	BRHS	England, Wales and Scotland	Men aged 40–59 years	Town	322.2	24	4.08	0.0127	0.0000164	0.0125	0.00131	1.42
5 year incidence of non-fatal ischaemic heart disease	BRHS	England, Wales and Scotland	Men aged 40–59 years	Town	322.2	24	7.42	0.0230	0.0000606	0.0224	0.00270	1.87
5 year incidence of fatal and non-fatal ischaemic heart disease	BRHS	England	Men aged 40–59 years	Town	322.2	24	11.5	0.036	0.0001	0.034	0.00293	1.94

[a] *Towns in the BRHS were not randomly sampled*

TABLE 28 *Cancer mortality and incidence data*

Variable	Source of data	Setting	Unit type	Cluster type	Cluster size	Number of clusters	Average number of events per cluster	Incidence rate	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Lung cancer mortality	PHCDS	England, 1995	Men aged < 75 years	DHA	216,947	105	113.1	0.000521	1.5×10^{-8}	0.000520	0.0000284	7.16
Lung cancer mortality	PHCDS	England, 1995	Women aged < 75 years	DHA	215,232	105	59.1	0.000274	7.0×10^{-9}	0.000274	0.0000271	6.83
Lung cancer mortality	PHCDS	England, 1995	All aged < 75 years	DHA	432,180	105	172.2	0.000398	1.0×10^{-8}	0.000398	0.0000261	12.3
Prostate cancer incidence	TCR	England, 1992	Men aged 65–74 years	DHA	18,484	27	48	0.00256	1.8×10^{-7}	0.00256	0.000072	2.33
Colon cancer incidence	TCR	England, 1992	Men aged 45–64 years	DHA	52,031	27	31	0.000593	1.3×10^{-9}	0.000592	0.0000223	2.16
Colon cancer incidence	TCR	England, 1992	Women aged 45–64 years	DHA	53,420	27	26	0.00482	7.0×10^{-9}	0.000481	0.0000141	1.76
Breast cancer incidence	TCR	England, 1992	Women aged 50–64 years	DHA	36,983	27	94	0.00252	4×10^{-8}	0.00251	0.000016	1.59

PHCDS, Public Health and Common Data Set; TCR, Thames Cancer Registry

TABLE 29 *Respiratory 1. Prevalence of asthma and bronchitis at town or local authority level*

Variable	Source of data	Setting	Unit type	Cluster type	Cluster size	Number of clusters	Average number of cases per cluster	Overall pro-portion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Wheeze	Burney (1991)[217]	England, 1986	Men aged 20–44 years	Local authority[a]	3474	21	411	0.117	0.0000768	0.103	0.000745	3.59
Night-time breathlessness	Burney (1991)[217]	England, 1986	Men aged 20–44 years	Local authority[a]	3474	21	135	0.0383	0.0000111	0.0368	0.000302	2.05
Self-reported asthma	Burney (1991)[217]	England, 1986	Men aged 20–44 years	Local authority[a]	3474	21	121	0.0343	0.000430	0.0331	0.00130	45.51
Phlegm (symptom)	BRHS[215]	England, Wales and Scotland	Men aged 40–59 years	Town[b]	322.2	24	51.4	0.160	0.00222	0.132	0.0165	6.30
Wheeze (symptom)	BRHS[215]	England, Wales and Scotland	Men aged 40–59 years	Town[b]	322.2	24	60.6	0.188	0.00381	0.149	0.0249	9.00
Breathlessness (symptom)	BRHS[215]	England, Wales and Scotland	Men aged 40–59 years	Town[b]	322.2	24	55.3	0.172	0.00211	0.140	0.0148	5.75
Doctor-diagnosed bronchitis	BRHS[215]	England, Wales and Scotland	Men aged 40–59 years	Town[b]	322.2	24	58.1	0.180	0.00236	0.146	0.0159	6.11
Doctor-diagnosed asthma	BRHS[215]	England, Wales and Scotland	Men aged 40–59 years	Town[b]	322.2	24	11.96	0.0371	0.000239	0.0355	0.00670	3.15
Having asthma not on inhaled β_2 agonists	Burney (1991)[217]	England, 1986	Men aged 20–44 years	Local authority[a]	132	21	56	0.415	0.00149	0.242	0.00614	1.81
Having asthma not on inhaled steroids	Burney (1991)[217]	England, 1986	Men aged 20–44 years	Local authority[a]	132	21	106	0.785	0.000181	0.169	0.00107	1.14

[a] *Local authorities were not randomly sampled*
[b] *Towns were not randomly sampled*

TABLE 30 *Respiratory 2. Symptoms, diagnoses and treatment at general practice level 1*

Variable	Source of data	Setting	Unit type	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall proportion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Chest wheezy or whistling	Premaratne (1997)[218]	London, UK, 1996	Patients	General practice	296.5	42	88.6	0.306	0.00130	0.211	0.0062	2.79
Woken with tight chest	Premaratne (1997)[218]	London, UK 1996	Patients	General practice	295.4	42	68.0	0.236	0.00111	0.179	0.0063	2.85
Woken short of breath	Premaratne (1997)[218]	London, UK, 1996	Patients	General practice	296.0	42	34.0	0.118	0.00063	0.104	0.0062	2.84
Woken with cough	Premaratne (1997)[218]	London, UK, 1996	Patients	General practice	295.3	42	96.4	0.334	0.00068	0.222	0.0031	1.91
Asthma attack in last 12 months	Premaratne (1997)[218]	London, UK, 1996	Patients	General practice	296.4	42	28.9	0.100	0.00042	0.090	0.0048	2.42
Taking medication for asthma	Premaratne (1997)[218]	London, UK, 1996	Patients	General practice	295.6	42	40.2	0.139	0.00108	0.119	0.0093	3.74
Patients with 'asthma'	Premaratne (1997)[218]	London, UK, 1996	Patients	General practice	299.1	42	56.5	0.240	0.00147	0.155	0.0096	3.86
Nasal allergies including hayfever	Premaratne (1997)[218]	London, UK, 1996	Patients	General practice	295.5	42	96.1	0.500	0.00030	0.222	0.0014	1.41
Ever smoked	Premaratne (1997)[218]	London, UK, 1996	Patients	General practice	297.2	42	126.2	0.435	0.00177	0.244	0.0074	3.19

TABLE 31 *Respiratory 3. Symptoms, diagnoses and treatment at general practice level 2*

Variable	Source of data	Setting	Unit type	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall proportion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Smoked in last month	Premaratne (1997)[218]	London, UK, 1996	Patients	General practice	297.0	42	103.4	0.357	0.00226	0.227	0.0101	3.99
Have steroid inhaler	Premaratne (1997)[218]	London, UK, 1996	Asthmatic patients	General practice	58.9	42	30.7	0.535	0.0000	0.249	0.0000	1.00
Have peak flow meter	Premaratne (1997)[218]	London, UK, 1996	Asthmatic patients	General practice	58.9	42	15.8	0.275	0.0105	0.189	0.0538	4.12
Have steroid tablets	Premaratne (1997)[218]	London, UK, 1996	Asthmatic patients	General practice	59.8	42	4.8	0.083	0.000307	0.076	0.0041	1.24
Asthma education	Premaratne (1997)[218]	London, UK, 1996	Asthmatic patients	General practice	58.6	42	23.5	0.410	0.00253	0.239	0.0107	1.62
Asthma-related quality of life (square root)	Premaratne (1997)[218]	London, UK, 1996	Asthmatic patients	General practice	61.1	42	–-	1.51	0.00394	0.407	0.0098	1.59

TABLE 32 *Health service activity 1. Health authority level data from health service indicators (HSI) 1*

Variable	Source of data	Setting	Unit type	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall pro-portion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
DP59: fertility rate	NHS HSI	England, 1994–1995	Women aged 15–44 years	DHA	91,531	111	5666	0.0619	0.000022	0.058	0.000384	36.1
HA55: cataract surgery operations, age 75+ years	NHS HSI	England, 1994–1995	Persons aged 75+ years	DHA	30,080	111	756	0.025	0.000049	0.024	0.0020	61.1
HA53: emergency admission with stroke, age 75+ years	NHS HSI	England, 1994–1995	Persons aged 75+ years	DHA	30,080	111	405	0.013	0.000017	0.0013	0.0129	389.0
HA54: hip replacement operations, age 75+ years	NHS HIS	England, 1994–1995	Persons aged 75+ years	DHA	30,080	111	93	0.003	0.00000048	0.003	0.00015	5.50
HA56: first psychiatric admissions, age 75+ years	NHS HSI	England, 1994–1995	Persons aged 75+ years	DHA	30,080	111	56	0.002	0.0000057	0.0018	0.0031	94.20

TABLE 33 *Health service activity 2. Health authority level data from HSI 2*

Variable	Source of data	Setting	Unit type	Cluster type	Cluster size	Number of clusters	Average number of cases per cluster	Overall pro-portion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
MT40: number of home deliveries	HSI 1994–1995	England	Deliveries	DHA	4186	104	33	0.0079	0.000169	0.0077	0.0216	91.44
MT67: proportion of Caesarean sections (non-elective)	HSI 1994–1995	England	Deliveries	DHA	4980	111	301	0.0603	0.001	0.0556	0.0178	89.56
MT68: proportion of forceps/Ventouse deliveries	HSI 1994–1995	England	Deliveries	DHA	4980	111	356	0.0714	0.00174	0.0646	0.0264	131.8
MT63: proportion of unintended home deliveries	HSI 1994–1995	England	Home deliveries	DHA	53	75	14	0.248	0.040	0.149	0.211	11.98
MT66: proportion of elective Caesarean sections	HSI 1994–1995	England	Deliveries	DHA	4997	110	215	0.043	0.00067	0.040	0.016	82.72
MT69: proportion of breech extraction or delivery	HSI 1994–1995	England	Deliveries	DHA	4980	111	31	0.071	0.0017	0.065	0.026	131.8
MT74: percentage of deliveries with first antenatal assessment at 12–19 weeks' gestation	HSI 1994–1995	England	Deliveries	DHA	5001	111	1515	0.302	0.062	0.149	0.295	1477
MT75: percentage of deliveries with first antenatal assessment at 20+ weeks' gestation or none	HSI 1994–1995	England	Deliveries	DHA	5001	111	2476	0.494	0.126	0.125	0.502	2511

TABLE 34 *Health service activity 3. Health authority level data from HSI 3*

Variable	Source of data	Setting	Unit type	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Average pro-portion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
HA58: emergency admissions for self-injury or poisoning	HSI 1994–1995	England	All people	DHA	438,143	111	407	0.00093	4.7×10^{-7}	0.00093	0.00051	225
HA57: patients statutorily detained	HSI 1994–1995	England	All people	DHA	439,200	104	151	0.00034	9.9×10^{-8}	0.00034	0.00029	128
IM44: proportion of girls immunised against rubella before 14 years of age	HSI 1994–1995	England, 1994–1995	Girls aged < 14 years	DHA	7914	116	5949	0.659	0.040	0.190	0.174	1378

TABLE 35 *Health service activity 4. Hospital level mortality data*

Variable	Source of data	Setting	Unit type	Cluster type	Cluster size	Number of clusters	Average number of cases per cluster	Overall pro-portion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Mortality in intensive care units	Intensive Care Society study[219]	Britain and Ireland	Intensive care unit patients	Intensive care unit	335	26	61	0.0179	0.0026	0.144	0.018	6.88
Mortality in hospitals	Intensive Care Society study[219]	Britain and Ireland	Intensive care unit patients	Hospital	335	26	94	0.0277	0.0036	0.197	0.018	6.93
Hospital mortality after acute upper gastrointestinal haemorrhage	Intensive Care Society study[219]	UK, 1993–1994	Patients aged 16 years and over	Hospital	74.92	74	10.7838	0.143396	0.00077	0.12210	0.00629	1.47

TABLE 36 *Health service activity 5. Data from the Royal College of Physicians (RCP) Wound Care Audit 1*[216]

Variable (proportion with)	Source of data	Setting	Unit type	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall pro-portion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Immobility	RCP	England and Wales, 1996	Patient with wound	Elderly care setting								
Audit 1					144.54	25	34.04	0.229	0.00956	0.168	0.0539	8.735
Audit 2					123.4	23	29.74	0.235	0.00863	0.172	0.0478	6.854
Incontinence of urine	RCP	England and Wales, 1996	Patient with wound	Elderly care setting								
Audit 1					143.49	25	28.36	0.193	0.00661	0.149	0.0420	7.04
Audit 2					122.05	23	24.78	0.198	0.00283	0.156	0.0178	3.153
Incontinence of faeces	RCP	England and Wales, 1996	Patient with wound	Elderly care setting								
Audit 1					144.23	25	21.24	0.143	0.0044	0.119	0.0358	6.122
Audit 2					122.46	23	18.83	0.150	0.00816	0.120	0.0637	8.736
Written care plan	RCP	England and Wales, 1996	Patient with wound	Elderly care setting								
Audit 1					143.36	25	138.6	0.941	0.00329	0.0522	0.0593	9.438
Audit 2					121.29	23	119.90	0.964	0.00129	0.0338	0.0367	5.420
No patient education recorded	RCP	England and Wales, 1996	Patient with wound	Elderly care setting								
Audit 1					140.87	25	36.48	0.252	0.0124	0.177	0.0655	10.16
Audit 2					117.97	23	28.00	0.231	0.0232	0.156	0.129	16.14
Leg ulcers > 5 cm	RCP	England and Wales, 1996	Patient with wound	Elderly care setting								
Audit 1					88.91	25	20.88	0.227	0.00129	0.174	0.00735	1.646
Audit 2					72.73	23	17.57	0.233	0.00416	0.175	0.0232	2.665

TABLE 37 *Health service activity 6. Data from the RCP Wound Care Audit 2*[216]

Variable (proportion with)	Source of data	Setting	Unit type	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall proportion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Pressure sore stage III or IV	RCP	England and Wales, 1996	Patient with wound	Elderly care setting								
Audit 1					49.21	25	15.92	0.317	0.00176	0.215	0.008	1.39
Audit 2					41.65	23	13.00	0.306	0.00237	0.210	0.0111	1.45
Wound not painful	RCP	England and Wales, 1996	Patient with wound	Elderly care setting								
Audit 1					95.70	25	17.04	0.172	0.000393	0.142	0.00276	1.261
Audit 2					90.63	22	15.64	0.167	0.000585	0.139	0.00420	1.376
Very satisfied	RCP	England and Wales, 1996	Patient with wound	Elderly care setting								
Audit 1					96.27	25	65.7	0.659	0.00949	0.216	0.042	5.009
Audit 2					90.48	22	56.9	0.609	0.00719	0.232	0.0301	3.694

TABLE 38 *Health service activity 7. Hospital level data from the HSI 1*

Variable	Source of data	Setting	Unit type	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall proportion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
CT02: percentage immediately assessed in accident and emergency department	HSI	England, 1996	Accident and emergency department attendance	Acute unit (hospital)	10,535.2	265	9757	0.925	0.00463	0.0649	0.0666	702.6
CT05: percentage of finished consultant episodes done as day cases	HSI	England, 1996	Finished consultant episode by selected speciality	Acute unit (hospital)								
C46: inguinal hernia					333.0	236	96	0.289	0.0259	0.180	0.129	43.8
C47: arthroscopy, knee					320.5	226	204	0.636	0.0381	0.194	0.164	53.5
C48: cataract extraction					1005.3	154	369	0.366	0.0546	0.178	0.235	236.9
C49: laparoscopic sterilisation					241.5	226	177	0.731	0.0290	0.168	0.147	36.5

TABLE 39 *Health service activity 8. Hospital level data from the HSI 2*

Variable	Source of data	Setting	Unit type	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall proportion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
MT46: percentage of deliveries by elective Caesarean section	HSI	England, 1996	Delivery	Acute unit (hospital)	2728.8	206	117.5	0.043	0.000890	0.0403	0.0216	59.95
MT47: percentage of deliveries by Caesarean section (other)	HSI	England, 1996	Delivery	Acute unit (hospital)	2728.8	206	165.0	0.0604	0.00142	0.0553	0.0251	69.43
MT48: percentage of deliveries by forceps, Ventouse or vacuum extraction	HSI	England, 1996	Delivery	Acute unit (hospital)	2728.8	206	195.6	0.0716	0.00231	0.0642	0.0348	95.90
MT49: percentage of deliveries by breech extraction or delivery	HSI	England, 1996	Delivery	Acute unit (hospital)	2722.9	205	15.5	0.00568	0.0000218	0.00563	0.00386	11.50

TABLE 40 *Health service activity 9. Family health service authority (FHSA) level data from HSI 1*

Variable	Source of data	Setting	Unit type	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall proportion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
CT10: percentage of general practices with practice charter	HSI	England, 1996	General practice	FHSA	99.8	90	64	0.639	0.0372	0.194	0.161	16.9
XM34: percentage of GPs practising single handed	HSI	England, 1996	GP	FHSA	293.9	90	31.4	0.106	0.00475	0.0900	0.0500	15.6
XM35: percentage of GPs aged > 65 years	HSI	England, 1996	GP	FHSA	293.9	90	4.5	0.0151	0.000123	0.0148	0.00827	3.42
XM24: percentage of GPs with list size > 2500	HSI	England, 1996	GP	FHSA	293.9	90	23.6	0.0800	0.00358	0.0701	0.0486	15.2
XM39: percentage of GPs with deputising service contract	HSI	England, 1996	GP	FHSA	293.9	90	174.9	0.592	0.0864	0.156	0.356	105.2
XM51: percentage of GPs achieving higher childhood immunisation targets	HSI	England, 1996	GP	FHSA	295.1	89	241.7	0.816	0.0243	0.127	0.161	48.4
XM52: percentage of GPs achieving lower childhood immunisation targets	HSI	England, 1996	GP	FHSA	295.1	89	35.5	0.120	0.00716	0.0984	0.0678	20.94
XM55: percentage of GPs achieving higher rate for preschool booster targets	HSI	England, 1996	GP	FHSA	295.1	89	241.8	0.816	0.0294	0.121	0.195	58.31

TABLE 41 *Health service activity 10. FHSA level data from HSI 2*

Variable	Source of data	Setting	Unit type	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall proportion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
XM56: percentage of GPs achieving lower rate preschool booster targets	HSI	England, 1996	GP	FHSA	295.1	89	34.2	0.115	0.00703	0.0951	0.0688	21.25
XM61: percentage of GPs achieving higher rate for cervical cytology	HSI	England, 1996	GP	FHSA	295.1	89	265.1	0.895	0.0276	0.0670	0.292	86.86
XM62: percentage of GPs achieving lower rate for cervical cytology	HSI	England, 1996	GP	FHSA	295.1	89	27.02	0.0912	0.0180	0.0652	0.216	64.56
XM71: percentage of GPs on child health surveillance list	HSI	England, 1996	GP	FHSA	294.0	90	276.7	0.937	0.00309	0.0556	0.0527	16.45
XM94: percentage of practices without a nurse	HSI	England, 1996	General practice	FHSA	100.8	86	7.1	0.0698	0.00360	0.0614	0.0554	6.53
XM98: percentage of practices below minimum standards	HSI	England, 1996	General practice	FHSA	100.4	82	6.66	0.0661	0.0191	0.0429	0.308	31.6
XM72: percentage of children for whom child health surveillance provided by GPs	HSI	England, 1996	Children aged < 5 years	FHSA	34,907.1	90	25,856	0.738	0.0190	0.175	0.0978	3414.7
XA48: patients removed from lists at doctor's request	HSI	England, 1996	Patients registered with GPs	FHSA	559,766	80	905	0.00161	0.00000266	0.00160	0.00165	928.0
XE42: sight tests in children	HSI	England, 1996	Children aged < 16 years	FHSA	112,987	90	25,100	0.221	0.0394	0.134	0.227	25,683.2

TABLE 42 *Other national health strategy target indicators*

Variable	Source of data	Setting	Unit type	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall pro-portion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Suicide and undetermined injury	PHCDS	1995	Men	DHA	228,321	105	34.0	0.000149	1.0×10^{-9}	0.000149	0.00000528	2.21
Suicide and undetermined injury	PHCDS	1995	Women	DHA	236,771	105	11.6	0.0000490	$< 10^{-9}$	0.0000490	0.00000124	1.29
Suicide and undetermined injury	PHCDS	1995	All	DHA	465,092	105	45.6	0.0000980	$< 10^{-9}$	0.0000980	0.00000309	2.44
Suicide	PHCDS	1995	Men	DHA	228,321	105	24.4	0.000107	$< 10^{-9}$	0.000107	0.00000461	2.05
Suicide	PHCDS	1995	Women	DHA	236,771	105	6.78	0.0000286	$< 10^{-9}$	0.0000286	0.000000359	1.09
Suicide	PHCDS	1995	All	DHA	465,092	105	31.2	0.0000670	$< 10^{-9}$	0.0000670	0.00000306	2.42
Fatal accidents	PHCDS	1995	Men aged < 15 years	DHA	46,030	105	2.67	0.0000579	$< 10^{-9}$	0.0000578	0.00000293	1.13
Fatal accidents	PHCDS	1995	Women aged < 15 years	DHA	43,696	105	1.30	0.0000298	$< 10^{-9}$	0.0000298	0.00000757	1.33
Fatal accidents	PHCDS	1995	All aged < 15 years	DHA	89,726	105	3.97	0.0000442	$< 10^{-9}$	0.0000442	0.00000624	1.56
Fatal accidents	PHCDS	1995	Men aged 15–24 years	DHA	30,052	105	8.31	0.000276	6.0×10^{-9}	0.000276	0.0000216	1.65
Fatal accidents	PHCDS	1995	Women aged 15–24 years	DHA	28,506	105	2.05	0.0000717	$< 10^{-9}$	0.0000717	0.000000531	1.02
Fatal accidents	PHCDS	1995	All aged 15–24 years	DHA	58,558	105	10.4	0.000177	2.0×10^{-9}	0.000177	0.00000964	1.56
Fatal accidents	PHCDS	1995	Men aged 65–84 years	DHA	27,740	105	12.2	0.000439	4.0×10^{-9}	0.000439	0.00000923	1.26
Fatal accidents	PHCDS	1995	Women aged 65–84 years	DHA	37,279	105	13.5	0.000362	7.0×10^{-9}	0.000362	0.0000181	1.67
Fatal accidents	PHCDS	1995	All aged 65–84 years	DHA	65,019	105	25.7	0.000398	5.0×10^{-9}	0.000395	0.0000134	1.87

TABLE 43 *Data for other Public Health Common data set indicators*

Variable	Source of data	Setting	Unit type	Cluster type	Average cluster size	Number of clusters	Average number of cases per cluster	Overall pro-portion	Variance component – between cluster	Variance component – within cluster	Intraclass correlation coefficient	Design effect
Number of still births per total births	PHCDS	1995	All births in 1995	DHA	5863	105	32.4	0.00552	8.5×10^{-7}	0.00549	0.000154	1.90
Infant mortality < 7 days of age	PHCDS	1995	Live births in 1995	DHA	5831	105	19.0	0.00325	6.1×10^{-7}	0.00324	0.000187	2.09
Infant mortality < 28 days of age	PHCDS	1995	Live births in 1995	DHA	5831	105	24.1	0.00413	9.6×10^{-7}	0.00411	0.000232	2.35
Infant mortality < 1 year of age	PHCDS	1995	Live births in 1995	DHA	5831	105	35.4	0.00606	1.4×10^{-6}	0.00602	0.000236	2.38
Mortality < 5 years of age	PHCDS	1995	All aged < 5 years	DHA	30,469	105	41.9	0.00137	6.7×10^{-8}	0.00137	0.0000489	2.49
Mortality < 15 years of age	PHCDS	1995	All aged < 15 years	DHA	89,726	105	50.8	0.000566	1.1×10^{-8}	0.000565	0.0000190	2.70
Perinatal deaths	PHCDS	1995	Live births in 1995	DHA	5831	105	51.4	0.00880	2.4×10^{-6}	0.00872	0.000276	2.61
Postneonatal deaths	PHCDS	1995	Live births in 1995	DHA	5831	105	11.3	0.00193	1.8×10^{-7}	0.00193	0.0000924	1.54

Chapter 10

Concluding remarks

Our aim in carrying out this review was to provide concise recommendations which would be easily applied by those involved in healthcare evaluation. Over the last few years a considerable amount of research has been carried out to address the methodological problems which are encountered in area- and organisation-based evaluations. These methods are now sufficiently accessible to allow implementation in the context of much healthcare evaluation. Further reviews are provided by Murray[40] and Donner and Klar.[11] The reader is also referred to Donner[65] for some recent research recommendations. For example, further work is needed: to aid the design of quasi-experimental cluster-based studies; to provide intraclass correlations and components of variance for a range of outcomes and different types of organisational clustering; to provide analytical methods for different types of data including ordinal and survival data; and to permit meta-analyses of the results of cluster-based studies.

Acknowledgements

The authors would like to thank the following who contributed to the work of the review: Professor Allan Donner, for advising on the conduct of the review and for commenting on the final report; Dr Peter Whincup, for providing access to unpublished data from the British Regional Heart Study; Dr Edward Dickinson and Ms Joy Windsor, for providing unpublished data from the Royal College of Physicians Wound Care Audit; and the referees, for their perseverance in reading the report and the quality of their comments.

Data from the Health Survey for England 1994 is Crown copyright. It has been made available by the Office for National Statistics through the Data Archive and has been used by permission. Neither the Office for National Statistics nor the Data Archive bear any responsibility for the analysis or interpretation of the data reported here.

References

1. Farquhar JW. The community-based model of life style intervention trials. *Am J Epidemiol* 1978;**108**:103–11.
2. Sherwin R. Controlled trials of the diet-heart hypothesis: some comments on the experimental unit. *Am J Epidemiol* 1978;**108**:92–9.
3. Puska P. Intervention and experimental studies. In: Holland WW, Detels R, Knox G, editors. Oxford textbook of public health. Oxford: Oxford Medical Publications, 1991:177–87.
4. Diwan VK, Eriksson B, Sterky G, Tomson G. Randomisation by group in studying the effect of drug information in primary care. *Int J Epidemiol* 1992;**21**:124–30.
5. Puska P, Salonen JT, Tuomilehto J, Nissinen A, Kottke TE. Evaluating community-based preventive cardiovascular programs: problems and experiences from the North Karelia project. *J Community Health* 1983;**9**:49–64.
6. Gail MH, Mark SD, Carroll RJ, Green SB, Pee D. On design considerations and randomisation-based inference for community intervention trials. *Stat Med* 1996;**15**:1069–92.
7. Buck C, Donner A. The design of controlled experiments in the evaluation of non-therapeutic interventions. *J Chron Dis* 1982;**35**:531–8.
8. Kirkwood BR, Morrow RH. Community-based intervention trials. *J Biosocial Sci Suppl* 1989;**10**:79–86.
9. 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.
10. Koepsell TD, Wagner EH, Cheadle AC, *et al.* Selected methodological issues in evaluating community-based health promotion and disease prevention programs. *Annu Rev Public Health* 1992;**13**:31–57.
11. Donner A, Klar N. Cluster randomisation trials in epidemiology: theory and application. *J Stat Plan Inference* 1994;**42**:37–56.
12. Simpson JM, Klar N, Donner A. Accounting for cluster randomisation: a review of primary prevention trials, 1990 through 1993. *Am J Public Health* 1995;**85**:1378–83.
13. Alexander F, Roberts MM, Lutz W, Hepburn W. Randomisation by cluster and the problem of social class bias. *J Epidemiol Community Health* 1989;**43**:29–36.
14. Simon R. Composite randomisation designs for clinical trials. *Biometrics* 1981;**37**:723–31.
15. Green SB, Corle DK, Gail MH, *et al.* Interplay between design and analysis for behavioural intervention trials with community as the unit of randomisation. *Am J Epidemiol* 1995;**142**:587–93.
16. Hulley SB. Symposium on coronary heart disease prevention trials: design issues in testing life style intervention. *Am J Epidemiol* 1978;**108**:85–6.
17. Kramer MS, Shapiro SH. Scientific challenges in the application of randomised trials. *JAMA* 1984;**252**:2739–45.
18. Syme SL. Life style intervention in clinic-based trials. *Am J Epidemiol* 1978;**108**:87–91.
19. Hsieh FY. Sample size formulae for intervention studies with the cluster as unit of randomisation. *Stat Med* 1988;**7**:1195–201.
20. Burmeister LF. Cluster sampling in hospital surveillance. *Infect Control Hosp Epidemiol* 1989;**10**:573–5.
21. Duffy SW, South MC, Day NE. Cluster randomisation in large public health trials: the importance of antecedent data. *Stat Med* 1992;**11**:307–16.
22. Whiting-O'Keefe QE, Henke C, Simborg DW. Choosing the correct unit of analysis in medical care experiments. *Med Care* 1984;**22**:1101–14.
23. McKinlay JB. More appropriate evaluation methods for community-level health interventions. *Evaluation Rev* 1996;**20**:237–43.
24. Fortmann SP, Flora JA, Winkleby MA, *et al.* Community intervention trials: reflections on the Stanford Five-City Project experience. *Am J Epidemiol* 1995;**142**:576–86.
25. Nicholl J, Turner J. Effectiveness of a regional trauma system in reducing mortality from major trauma: before and after study. *BMJ* 1997;**315**:1349–54.
26. Dwyer JH, MacKinnon DP, Pentz MA, *et al.* Estimating intervention effects in longitudinal studies. *Am J Epidemiol* 1989;**130**:781–95.
27. Williams PT, Fortmann SP, Farquhar JW, Varady A, Mellen S. A comparison of statistical methods for evaluating risk factor changes in community-based studies: an example from the Stanford Three-Community Study. *J Chron Dis* 1981;**34**:565–71.

28. Kelder SH, Jacobs DR, Jr, Jeffery RW, McGovern PG, Forster JL. The worksite component of variance: design effects and the Healthy Worker Project. *Health Educ Res* 1993;**8**:555–66.

29. Donner A. Approaches to sample size estimation in the design of clinical trials: a review. *Stat Med* 1984;**3**:199–214.

30. Oxman AD. Checklists for review articles. In: Chalmers II, Altman DG, editors. Systematic reviews. London: BMJ, 1995:75–85.

31. Hutton JL, Ashcroft RE. What does 'systematic' mean for reviews of methodology? In: Black N, Brazier J, Fitzpatrick R, Reeves B, editors. Methods for health services research. A state of the art guide. London: BMJ, 1998:249–54.

32. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. In: Chalmers I, Altman DG, editors. Systematic reviews. London: BMJ, 1995:17–36.

33. Slavin RE. Best-evidence synthesis: an intelligent alternative to meta-analysis. *J Clin Epidemiol* 1995;**48**:9–18.

34. 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.

35. Hayes R, Mosha F, Nicoll A, *et al.* A community trial of the impact of improved sexually transmitted disease treatment on the HIV epidemic in rural Tanzania: 1. Design. *AIDS* 1995;**9**:919–26.

36. Salonen JT, Kottke TE, Jacobs DR, Jr., Hannan PJ. Analysis of community-based cardiovascular disease prevention studies – evaluation issues in the North Karelia Project and the Minnesota Heart Health Program. *Int J Epidemiol* 1986;**15**:176–82.

37. Last JM. A dictionary of epidemiology. Oxford: Oxford University Press, 1988.

38. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health* 1998;**52**:377–84.

39. Meinert CL. Clinical trials. Design, conduct and analysis. Oxford: Oxford University Press, 1986.

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

41. Campbell DT, Stanley JC. Experimental and quasi-experimental designs for research. Chicago: Rand McNally College Publishing, 1963.

42. Cook TD, Campbell DT. Quasi-experimentation. Design and analysis issues for field settings. Chicago: Rand McNally College Publishing, 1979.

43. Goodwin N, Mays N, McLeod H, Malbon G, Raftery J, on behalf of the Total Purchasing National Evaluation Team. Evaluation of total purchasing pilots in England and Scotland and implications for primary care groups in England: personal interviews and analysis of routine data. *BMJ* 1998;**317**:256–9.

44. Dowse GK, Gareebo H, Alberti KGMM, *et al.* Changes in population cholesterol concentrations and other cardiovascular risk factor levels after five years of the non-communicable disease intervention programme in Mauritius. *BMJ* 1995;**311**:1208–12.

45. Tudor-Smith C, Nutbeam D, Moore L, Catford J. Effects of Heartbeat Wales programme over five years on behavioural risks for cardiovascular disease: quasi-experimental comparison of results from Wales and a matched reference area. *BMJ* 1998;**316**:818–22.

46. McCarthy M. The benefit of seat belt legislation in the United Kingdom. *J Epidemiol Community Health* 1989;**43**:218–22.

47. Somerville SM, Rona RJ, Chinn S, Qureshi S. Family Credit and uptake of school meals in primary school. *J Public Health Med* 1996;**18**:98–106.

48. Pan XR, Li GW, Hu YH, *et al.* Effect of diet and exercise in preventing NIDDM in people with impaired glucose tolerance. The Da Qing IGT and Diabetes Study. *Diabetes Care* 1997;**20**:537–44.

49. Last JM. A dictionary of epidemiology. Oxford: Oxford University Press, 1988.

50. Freedman LS, Green SB, Byar DP. Assessing the gain in efficiency due to matching in a community intervention study. *Stat Med* 1990;**9**:943–52.

51. Klar N, Donner A. The merits of matching: a cautionary tale. *Stat Med* 1997;**16**:1753–64.

52. Koepsell TD, Diehr PH, Cheadle A, Kristal A. Invited commentary: symposium on community intervention trials. *Am J Epidemiol* 1995;**142**:594–9.

53. Murray DM. Design and analysis of community trials: lessons from the Minnesota Heart Health Program. *Am J Epidemiol* 1995;**142**:569–75.

54. Jacobs DR, Jr., Luepker RV, Mittelmark M. Community-wide prevention strategies: evaluation design of the Minnesota Heart Health Program. *J Chron Dis* 1986;**39**:775–88.

55. Martin DC, Diehr P, Perrin EB, Koepsell TD. The effect of matching on the power of randomised community intervention studies. *Stat Med* 1993;**12**:329–38.

56. Gail MH, Byar DP, Pechacek TF, Corle DK. Aspects of statistical design for the Community Intervention Trial for Smoking Cessation (COMMIT). *Controlled Clin Trials* 1992;**13**:6–21.

57. Thompson SG, Pyke SM, Hardy RJ. The design and analysis of paired cluster randomised trials: an application of meta-analysis techniques. *Stat Med* 1997;**16**:2063–79.

58. Donner A, Donald A. Analysis of data arising from a stratified design with the cluster as unit of randomisation. *Stat Med* 1987;**6**:43–52.

59. Donner A, Klar N. Confidence interval construction for effect measures arising from cluster randomisation trials. *J Clin Epidemiol* 1993;**46**:123–31.

60. Shipley MJ, Smith PG, Dramaix M. Calculation of power for matched pair studies when randomisation is by group. *Int J Epidemiol* 1989;**18**:457–61.

61. Thompson SG, Pyke SD, Hardy RJ. The design and analysis of paired cluster randomised trials: an application of meta-analysis techniques. *Stat Med* 1997;**16**:2063–79.

62. Donner A, Hauck W. Estimation of a common odds ratio in case-control studies of familial aggregation. *Biometrics* 1988;**44**:369–78.

63. Family Heart Study Group. Randomised controlled trial evaluating cardiovascular screening and intervention in general practice: principal results of the British Family Heart Study. *BMJ* 1994;**308**:313–20.

64. Zucker DM, Lakatos E, Webber LS, *et al.* Statistical design of the Child and Adolescent Trial for Cardiovascular Health (CATCH): implications of cluster randomisation. *Controlled Clin Trials* 1995;**16**:96–118.

65. Donner A. Some aspects of the design and analysis of cluster randomisation trials. *Appl Stat* 1998;**47**:95–113.

66. Elliott TE, Murray DM, Oken MM, Johnson KM, Elliott BA, Post-White J. The Minnesota Cancer Pain Project: design, methods, and education strategies. *J Cancer Educ* 1995;**10**:102–12.

67. 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.

68. Chinn S. Mixed longitudinal studies: their efficiency for the estimation of trends over time. *Ann Human Biol* 1988;**15**:443–54.

69. Koepsell TD, Martin DC, Diehr PH, *et al.* Data analysis and sample size issues in evaluations of community-based health promotion and disease prevention programs: a mixed-model analysis of variance approach. *J Clin Epidemiol* 1991;**44**:701–13.

70. Murray DM, Hannan PJ. Planning for the appropriate analysis in school-based drug-use prevention studies. *J Consult Clin Psychol* 1990;**58**:458–68.

71. 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.

72. McKinlay SM. Cost-efficient designs of cluster unit trials. *Prev Med* 1994;**23**:606–11.

73. Gillum RF, Williams PT, Sondik E. Some considerations for the planning of total-community prevention trials – when is sample size adequate? *J Community Health* 1980;**5**:270–8.

74. Windsor RA, Kronenfeld JJ, Ory MG, Kilgo JS. Method and design issues in evaluation of community health education programs: a case study in breast and cervical cancer. *Health Educ Q* 1980;**7**:203–18.

75. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;**312**:1215–18.

76. Weed DL. On the use of causal criteria. *Int J Epidemiol* 1997;**26**:1137–41.

77. Griffiths C, Sturdy P, Naish J, Omar R, Dolan S, Feder G. Hospital admissions for asthma in East London: associations with characteristics of local general practices, prescribing and population. *BMJ* 1997;**314**:482–6.

78. Gulliford MC, Petruckevitch A, Burney PG. Survival with bladder cancer, evaluation of delay in treatment, type of surgeon and modality of treatment. *BMJ* 1991;**303**:437–40.

79. 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, editors. Health services research methods. A guide to best practice. London: BMJ, 1998:98–107.

80. Klar N, Gyorkos T, Donner A. Cluster randomisation trials in tropical medicine: a case study. *Trans R Soc Trop Med Hyg* 1995;**89**:454–9.

81. Zucker DM. An analysis of variance pitfall: the fixed effects analysis in a nested design. *Educ Psychol Measure* 1990;**50**:731–8.

82. Snedecor GW, Cochran WG. Statistical methods. Ames: Iowa State University Press, 1967.

83. Donner A. The analysis of intraclass correlation in multiple samples. *Ann Hum Genet* 1985;**49**:75–82.

84. Kish L. Survey sampling. London: Wiley, 1965:148–81.

85. Moser CA, Kalton G. Survey methods in social investigation. Aldershot: Dartmouth Publishing, 1993:61–78.

86. Donner A. An empirical study of cluster randomisation. *Int J Epidemiol* 1982;**11**:283–6.

87. McCulloch CE. Random effects models for binary data applied to environmental/ecological studies. Biometrics Unit, Cornell University, Ithaca, 1991.

88. Katz J, Carey VJ, Zeger SL, Sommer A. Estimation of design effects and diarrhea clustering within households and villages. *Am J Epidemiol* 1993;**138**:994–1006.

89. Mickey RM, Goodwin GD. The magnitude and variability of design effects for community intervention studies. *Am J Epidemiol* 1993;**137**:9–18.

90. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Measure* 1973;**33**:613–19.

91. Fleiss JL. Statistical methods for rates and proportions. Chichester: Wiley, 1981;211–36.

92. Katz J, Zeger SL. Estimation of design effects in cluster surveys. *Ann Epidemiol* 1994;**4**:295–301.

93. Carey VJ, Zeger SL, Diggle PJ. Modelling multivariate binary data with alternating logistic regressions. *Biometrika* 1993;**80**:517–26.

94. Muller R, Buttner P. A critical discussion of intraclass correlation coefficients. *Stat Med* 1994;**13**:2465–76.

95. Searle SR, Bradley RA, Hunter JS, Kendall DG, Watson GS, editors. Linear models. Chichester: Wiley, 1971.

96. Wald A. A note on the analysis of variance with unequal class frequencies. *Ann Math Stat* 1940;**11**:96–100.

97. Smith CAB. On the estimation of intraclass correlation. *Ann Hum Genet* 1956;**21**:363–73.

98. Donner A, Wells G. A comparison of confidence interval methods for the intraclass correlation coefficient. *Biometrics* 1986;**42**:401–12.

99. Minitab Statistical Software. Minitab reference manual. Release 8. PC version. Philadelphia: Minitab, 1991.

100. SAS Institute. SAS/STAT user guide. Version 6, vol 2. Cary, NC: SAS Institute, 1990:1127–34.

101. Gleason JR. Computing intraclass correlations and large ANOVAs. *Stata Tech Bull* 1997;**35**:25–31.

102. Feng Z, Grizzle JE. Correlated binomial variates: properties of estimator of intraclass correlation and its effect on sample size calculation. *Stat Med* 1992;**11**:1607–14.

103. 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 Program. *Epidemiology* 1994;**5**:88–95.

104. Mickey RM, Goodwin GD, Costanza MC. Estimation of the design effect in community intervention studies. *Stat Med* 1991;**10**:53–64.

105. Murray DM, Short B. Intraclass correlation among measures related to alcohol use by young adults: estimates, correlates and applications in intervention studies. *J Studies Alcohol* 1995;**56**:681–94.

106. Siddiqui O, Hedeker D, Flay BR, Hu FB. Intraclass correlation estimates in a school-based smoking prevention study – outcome and mediating variables, by sex and ethnicity. *Am J Epidemiol* 1996;**144**:425–33.

107. Raudenbush SW. Statistical analysis and optimal design for cluster randomised trials. *Psychol Methods* 1997;**3**:173–85.

108. Donner A, Birkett N, Buck C. Randomisation by cluster. Sample size requirements and analysis. *Am J Epidemiol* 1981;**114**:906–14.

109. Cornfield J. Randomisation by group: a formal analysis. *Am J Epidemiol* 1978;**108**:100–2.

110. Armitage P, Berry G. Statistical methods in medical research. Oxford: Blackwell Science, 1994:207–36.

111. Donner A. Sample size requirements for stratified cluster randomisation designs. *Stat Med* 1992;**11**:743–50.

112. Woolson RF, Bean JA, Rojas PB. Sample size for case–control studies using Cochran's statistic. *Biometrics* 1986;**42**:927–32.

113. Donner A, Klar N. Methods for comparing event rates in intervention studies when the unit of allocation is a cluster. *Am J Epidemiol* 1994;**140**:279–89 (discussion:300–1).

114. Donner A, Klar N. Statistical considerations in the design and analysis of community intervention trials. *J Clin Epidemiol* 1996;**49**:435–9.

115. Donner A, Donald A. The statistical analysis of multiple binary measurements. *J Clin Epidemiol* 1988;**41**:899–905.

116. Rao JN, Scott AJ. A simple method for the analysis of clustered binary data. *Biometrics* 1992;**48**:577–85.

117. Donner A, Hauck W. Estimation of a common odds ratio in paired-cluster randomisation designs. *Stat Med* 1989;**8**:599–607.

118. Donald A, Donner A. Adjustments to the Mantel–Haenszel chi-square statistic and odds ratio variance estimator when the data are clustered. *Stat Med* 1987;**6**:491–9.

119. Scott AJ, Holt B. The effect of two-stage sampling on ordinary least squares methods. *J Am Stat Assoc* 1982;**77**:848–54.

120. Goldstein H. Multilevel statistical models. London: Arnold, 1996:1–178.

121. Neuhaus JM, Segal MR. Design effects for binary regression models fitted to dependent data. *Stat Med* 1993;**12**:1259–68.

122. Aitkin M, Longford N. Statistical modelling issues in school effectiveness studies. *J R Stat Soc A* 1986;**149**:1–43.

123. Rice N, Leyland A. Multilevel models: applications to health data. *J Health Services Res Policy* 1996;**1**:154–64.

124. Singer JD. An intraclass correlation model for analysing multilevel data. *J Exp Educ* 1987;**55**:219–28.

125. Brown RL, Baumann LJ, Cameron L. A cautionary note regarding the use of single-level regression analyses of primary care intervention studies with hierarchically structured data. *Nurs Res* 1996; **45**:359–62.

126. Singleton RA, Straits BC, Straits MM. Approaches to social research. Oxford: Oxford University Press, 1993.

127. Luepker RV, Rastam L, Hannan PJ, *et al.* Community education for cardiovascular disease prevention – morbidity and mortality results from the Minnesota Heart Health Program. *Am J Epidemiol* 1996;**144**:351–62.

128. Bryk AS, Raudenbush SW. Hierarchical linear models. London: Sage, 1992.

129. Longford NT. Random coefficient models. Oxford: Oxford University Press, 1995.

130. Arnold CL. An introduction to hierarchical linear models. *Measure Eval Counselling Dev* 1992;**25**:58–90.

131. Paterson L, Goldstein H. New statistical methods for analysing social structures: an introduction to multilevel models. *Br Educ Res J* 1991;**17**:387–93.

132. Duncan C, Jones K, Moon G. Context, composition and heterogeneity: using multi-level models in health research. *Soc Sci Med* 1998;**46**:97–117.

133. Breslow E, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc* 1993;**88**:9–25.

134. Andersen PK. Survival analysis 1982–1991: the second decade of the proportional hazards regression model. *Stat Med* 1991;**10**:1931–41.

135. Guo SW, Lin DY. Regression analysis of multivariate grouped survival data. *Biometrics* 1994;**50**:632–9.

136. Pickles A, Crouchley R. A comparison of frailty models for multivariate survival data. *Stat Med* 1995;**14**:1447–61.

137. Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. *Biometrics* 1994;**50**:933–44.

138. Hedeker D, Gibbons RD. MIXOR: a computer program for mixed-effects ordinal regression analysis. *Comput Methods Programs Biomed* 1996;**49**:157–76.

139. Plewis I. Terminology and definition in multilevel models analysis. Multilevel modelling newsletter, vol 9, No 1. London, Institute of Education, University of London, 1997:2–4.

140. Ten Have TR, Landis JR, Hartzel J. Population-averaged and cluster-specific models for clustered ordinal response data. *Stat Med* 1996;**15**:2573–88.

141. Liang K, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986;**73**:13–22.

142. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986;**42**:121–30.

143. Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988;**44**:1049–60.

144. Diggle PJ, Liang KY, Zeger SL. Analysis of longitudinal data. Oxford: Oxford University Press, 1996.

145. Kreft IGG, de Leeuw J, van der Leeden R. Review of five multilevel analysis programs: BMDP-5V, GENMOD, HLM, ML3, VARCL. *Am Stat* 1994;**48**:324–35.

146. Rodriguez G, Goldman N. An assessment of estimation procedures for multilevel models with binary responses. *J R Stat Soc A* 1995;**158**:73–89.

147. Goldstein H. Multilevel models and generalised estimating equations. Multilevel modelling newsletter, vol 5, No 2. London, Institute of Education, University of London, 1993:2.

148. Goldstein H, Rasbash J. Improved approximations for multilevel models with binary responses. *J R Stat Soc A* 1996;**159**:505–13.

149. Goldstein H. Nonlinear multilevel models, with an application to discrete response data. *Biometrika* 1991;**78**:45–51.

150. de Leeuw J, Kreft IGG. Questioning multilevel models. *J Educ Behav Stat* 1995;**20**:171–89.

151. Reid SE, Simpson JM, Britt HC. Pap smears in general practice: a secondary analysis of the Australian Morbidity and Treatment Survey 1990 to 1991. *Aust NZ J Public Health* 1997;**21**:257–64.

152. Phibbs CS, Bronstein JM, Buxton E, Phibbs RH. The effects of patient volume and level of care at the hospital of birth on neonatal mortality. *JAMA* 1996;**276**:1054–9.

153. Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *J R Stat Soc A* 1996;**159**:385–443.

154. Stata Corporation. Stata statistical software. Release 5. College Station, TX: Stata Press, 1997.

155. Woodhouse G. Multilevel modelling applications. A guide for users of MLn. London: Institute of Education, University of London, 1998.

156. Grosskurth H, Mosha F, Todd J, *et al.* Impact of improved treatment of sexually transmitted diseases on HIV infection in rural Tanzania: randomised controlled trial. *Lancet* 1995;**346**:530–6.

157. Family Heart Study Group. The British Family Heart Study: its design and method, and prevalence of cardiovascular risk factors. *Br J Gen Pract* 1994;**44**:62–7.

158. van Teiljlingen ER, Friend JAR, Twine F. Evaluation of Grampian smokebusters. *J Public Health Med* 1996;**18**:143–51.

159. Snooks HA, Nicholl JP, Brazier JE, Lees Mlanga S. The costs and benefits of helicopter emergency ambulance services in England and Wales. *J Public Health Med* 1996;**18**:67–77.

160. Campbell H, MacDonald S. Evaluation of the women's drop-in service in Benarty, Fife. *J Public Health Med* 1996;**18**:143–51.

161. Perenboom RJM, Davidse W. Increasing the coverage of vaccination against influenza by general practitioners. *J Public Health Med* 1996;**18**:183–7.

162. Fulop NJ, Koffman J, Carson S, Robinson A, Pashley D, Coleman K. Use of acute psychiatric beditors: a point prevalence survey in North and South Thames regions. *J Public Health Med* 1996;**18**:207–16.

163. Barros H, Tavares M, Rodrigues T. Role of prenatal care in preterm birth and low birthweight in Portugal. *J Public Health Med* 1996;**18**:321–8.

164. Li PL, Logan S. The current state of screening in general practice. *J Public Health Med* 1996;**18**:350–6.

165. Mulholland C, Harding N, Bradley S, Stevenson M. Regional variations in the utilisation rate of vaginal and abdominal hysterectomies in the United Kingdom. *J Public Health Med* 1996;**18**:400–5.

166. Brekelmans CTM, Westers P, Faber JAJ, Peeters PHM, Collette HJA. Age specific sensitivity and sojourn time in a breast cancer screening programme in The Netherlands: a comparison of different methods. *J Epidemiol Community Health* 1996;**50**:68–71.

167. Sharp DJ, Peters TJ, Bartholemew J, Shaw A. Breast screening: a randomised controlled trial in UK general practice of three interventions designed to increase uptake. *J Epidemiol Community Health* 1996;**50**:72–6.

168. Horton Taylor D, McPherson K, Parbhoo S, Perry N. Response of women aged 65 to 74 to invitation for screening for breast cancer by mammography: a pilot study in London UK. *J Epidemiol Community Health* 1996;**50**:77–80.

169. Lindholm L, Reosen M, Weinehall L, Asplund K. Cost-effectiveness and equity of a community based cardiovascular disease prevention programme in Norsjo, Sweden. *J Epidemiol Community Health* 1996;**50**:190–5.

170. Wen SW, Simunovic M, Williams JI, Johnston KW, Naylor CD. Hospital volume, calendar age, and short term outcomes in patients undergoing repair of abdominal aortic aneurysm. *J Epidemiol Community Health* 1996;**50**:207–13.

171. Albert X, Bayo A, Alfonso JL, Cortina P, Corella D. The effectiveness of health systems in influencing avoidable mortality: a study in Valencia, Spain. *J Epidemiol Community Health* 1996;**50**:320–5.

172. Westerling R. Can regional variation in 'avoidable' mortality be explained by deaths outside hospital? A study from Sweden 1987–1990. *J Epidemiol Community Health* 1996;**50**:326–33.

173. Chenet L, McKee M. Challenges of monitoring use of secondary care at local level: a study based in London UK. *J Epidemiol Community Health* 1996;**50**:359–65.

174. Lang T, David A, Diatike B, Agay E, Viel JF, Flicoteaux B. Non urgent care in the hospital medical emergency department in France: how much and which health needs does it reflect? *J Epidemiol Community Health* 1996;**50**:456–62.

175. Perneger T, Etter JF, Raetzo MA, Schaller P, Stadler H. Comparison of patient satisfaction with ambulatory visits in competing health care delivery settings in Geneva, Switzerland. *J Epidemiol Community Health* 1996;**50**:463–8.

176. Clarke A, Rowe P, Black N. Does a shorter length of hospital stay affect the outcome and costs of hysterectomy in southern England? *J Epidemiol Community Health* 1996;**50**:545–50.

177. Ytterstad B. The Harstad injury prevention study: community based prevention of fall fractures in the elderly evaluated by means of a hospital based injury recording system in Norway. *J Epidemiol Community Health* 1996;**50**:551–8.

178. Otten JDM, van Dijck JAAM, Peer PGM, *et al.* Long term breast cancer screening in Nijmegen, the Netherlands: the nine rounds from 1975–1992. *J Epidemiol Community Health* 1996;**50**:353–8.

179. Rissanen P. Hospital and patient related characteristics determining length of hospital stay for hip and knee replacements. *Int J Technol Assess Health Care* 1996;**12**:325–35.

180. Favaretti C, Mariotto A. Time trends in utilisation of cardiac catheterisation procedures in Italy, 1983–1993. *Int J Technol Assess Health Care* 1996;**12**:518–23.

181. Wait SH, Allemand HM. The French breast cancer screening programme. Epidemiological and economic results of the first round of screening. *Eur J Public Health* 1996;**6**:43–8.

182. Taziaux P, Franck J, Ludovicy R, Albert A. A study of general practitioners' prescribing behaviour to the elderly in Wallonia, Belgium. *Eur J Public Health* 1996;**6**:49–57.

183. de la Fuente DO. Inappropriate hospitalisation. Reasons and determinants. *Eur J Public Health* 1996;**6**:126–32.

184. Fakhoury WKH, McCarthy M, Addington Hall JM. Which informal carers are most satisfied with services for dying cancer patients? *Eur J Public Health* 1996;**6**:181–7.

185. Normand SLT, Glickman ME, Sharma RG, McNeil BJ. Using admission characteristics to predict short-term mortality from myocardial infarction in elderly patients. *JAMA* 1996;**275**:1322–8.

186. Sisk JE, Gorman ME, Reisinger AL, Glied SA, DuMouchel WH, Hynes MM. Evaluation of Medicaid managed care. Satisfaction, access and use. *JAMA* 1996;**276**:50–5.

187. Ware JE, Bayliss MS, Rogers WH, Kosinski M, Tarlov AR. Differences in 4 year health outcomes for elderly and poor chronically ill patients treated in HMO and fee for service systems. *JAMA* 1996;**276**:1039–47.

188. Robinson JC. Decline in hospital utilisation and cost inflation under managed care in California. *JAMA* 1996;**276**:1060–4.

189. Iezzoni LI, Shwartz M, Ash AS, Hughes JS, Daley J, MacKiernan YD. Severity measurement methods and judging hospital death rates for pneumonia. *Med Care* 1996;**34**:11–28.

190. Smith CB, Goldman RL, Martin DC, *et al.* Overutilisation of acute care beds in Veteran's Affairs Hospitals. *Med Care* 1996;**34**:85–96.

191. Gordon HS, Rosenthal GE. Impact of inter-hospital transfers on outcomes in an academic medical centre. Implications for profiling hospital quality. *Med Care* 1996;**34**:295–309.

192. Roos LL, Wall RK, Romano PS, Roberecki S. Short term mortality after repair of hip fracture. Do Manitoba elderly do worse? *Med Care* 1996;**34**:310–26.

193. Martin BC, McMillan RP. The impact of implementing a more restrictive prescription limit on Medicaid recipients. *Med Care* 1996;**34**:686–701.

194. Sturm R, Meredith LS, Wells KB. Provider choice and continuity for the treatment of depression. *Med Care* 1996;**34**:723–34.

195. Young WW, Marks SM, Kohler SA, Hsu AY. Dissemination of clinical results. Mastectomy versus lumpectomy and radiation therapy. *Med Care* 1996;**34**:1003–17.

196. Gerdtham UG, Hertzman P, Jonsson B, Boman G. Impact of inhaled corticosteroids on acute asthma hospitalisation in Sweden 1978–1991. *Med Care* 1996;**34**:1188–98.

197. Hall JA, Roter DL, Milburn MA, Daltrey LH. Patients' health as a predictor of physician and patient behaviour in medical visits. *Med Care* 1996;**34**:1205–18.

198. Taylor FC, Ramsay ME, Renton A, Cohen H. Methods for managing the increased workload in anticoagulant clinics. *BMJ* 1996;**312**:286.

199. Sikorski J, Wilson J, Clement S, Das S, Smeeton N. A randomised controlled trial comparing two schedules of antenatal visits: the antenatal acre project. *BMJ* 1996;**312**:546–53.

200. Tucker JS, Hall MH, Howie PW, Reid ME, Barbour RS, Florey C du V. Should obstetricians see women with normal pregnancies? A multi-centre randomised controlled trial of routine antenatal care by general practitioners and midwives compared with shared care led by obstetricians. *BMJ* 1996;**312**:554–9.

201. MacGregor SH, Hamley JG, Dunbar JA, Dodd TRP, Cromarty JA. Evaluation of a primary care anticoagulant clinic managed by a pharmacist. *BMJ* 1996;**312**:560.

202. Glyn Jones E, Williams LA, Barry S, Kinnersley P. Waiting list management in general practice: a review of orthopaedic patients. *BMJ* 1996;**312**:887–88.

203. Giles GG, Armstrong BK, Burton RC, Staples MP. Has mortality from melanoma stopped rising in Australia? Analysis of trends in mortality between 1931 and 1994. *BMJ* 1996;**312**:1121–5.

204. Murphy AW, Bury G, Plunjett PK, *et al.* Randomised controlled trial of general practitioner versus medical care in an urban accident and emergency department. *BMJ* 1996;**312**:1135–42.

205. Kammerling RM, Kinnear A. The extent of the two tier service for fundholders. *BMJ* 1996;**312**:1399–401.

206. Wilson RPH, Hatcher J, Barton S, Walley T. Influences of practice characteristics on prescribing in fundholding and non-fundholding general practices. An observational study. *BMJ* 1996;**312**:595–9.

207. Stirland H, Husain OAN, Butler OAB, Russell KS. Cervical screening in the inner city: is the opportunistic approach still worthwhile? *BMJ* 1996;**312**:600.

208. MacDonald TM, McMahon AD, Reid IC, Fenton GW, McDevitt DG. Anti-depressant use in primary care: a record linkage study in Tayside Scotland. *BMJ* 1996;**312**:860–1.

209. Donoghue J, Tylee A, Wildgust H. Cross-sectional database analysis of anti-depressant prescribing in general practice in the United Kingdom 1993–1995. *BMJ* 1996;**312**:861–2.

210. Shelley M, Croft P, Chapman S, Pantin C. Is the ratio of inhaled corticosteroid to bronchodilator a good indicator of the quality of asthma prescribing? Cross-sectional study linking prescribing data to data on admissions. *BMJ* 1996;**313**:1124–6.

211. Yelin EH, Criswell LA, Feigenbaum PG. Health care utilisation and outcomes among persons with rheumatoid arthritis in fee for service and repaid group practice settings. *JAMA* 1996;**276**:1048–53.

212. Bingefors K, Isacson D, von Knorring L, Smedby B, Ekselius L, Kupper LL. Antidepressant-treated patients in ambulatory care. Long term use of non-psychotropic and psychotropic drugs. *Br J Psychiatry* 1996;**168**:292–8.

213. Murray DM, Rooney BL, Hannan PJ, *et al.* Intraclass correlation among common measures of adolescent smoking: estimates, correlates, and applications in smoking prevention studies. *Am J Epidemiol* 1994;**140**:1038–50.

214. Colhoun H, Prescott-Clarke P. Health survey for England 1994. London: HMSO, 1996.

215. Shaper AG, Pocock SJ, Walker M, Cohen NM, Wale CJ, Thomson AG. British Regional Heart Study: cardiovascular risk factors in middle-aged men in 24 towns. *BMJ* 1981;**283**:179–86.

216. Royal College of Physicians Research Unit. Older people's programme. National chronic wound audit. London: Royal College of Physicians, 1997.

217. Burney PGJ, Papacosta AO, Withey CH, Colley JR, Holland WW. Hospital admission rates and the prevalence of asthma symptoms in 20 local authority districts. *Thorax* 1991;**46**:574–9.

218. Premaratne UN, Sterne JAC, Webb J, Burney PGJ. A general practice based community intervention on the management of asthma. *Thorax* 1997;**52**:A18.

219. Rowan K, Kerr JH, Major E, McPherson K, Short A, Vessey MP. Intensive Care Society's APACHE II study in Britain and Ireland I. Variations in case mix of adult admissions to general intensive care units and impact on outcome. *BMJ* 1993;**307**:972–7.

220. Rockall TA, Logan RF, Devlin HB, Northfield TC. Variation in outcome after acute upper gastro-intestinal haemorrhage. The National Audit of Acute Upper Gastrointestinal Haemorrhage. *Lancet* 1995;**346**:346–50.

Health Technology Assessment panel membership

This report was identified as a priority by the Methodology Panel.

Acute Sector Panel

Current members

Chair:
Professor Francis H Creed, University of Manchester

Professor Clifford Bailey, University of Leeds

Ms Tracy Bury, Chartered Society of Physiotherapy

Professor Collette Clifford, University of Birmingham

Dr Katherine Darton, M.I.N.D.

Mr John Dunning, Papworth Hospital, Cambridge

Mr Jonathan Earnshaw, Gloucester Royal Hospital

Mr Leonard Fenwick, Freeman Group of Hospitals, Newcastle-upon-Tyne

Professor David Field, Leicester Royal Infirmary

Ms Grace Gibbs, West Middlesex University Hospital NHS Trust

Dr Neville Goodman, Southmead Hospital Services Trust, Bristol

Professor Mark P Haggard, MRC

Professor Robert Hawkins, University of Manchester

Dr Duncan Keeley, General Practitioner, Thame

Dr Rajan Madhok, East Riding Health Authority

Dr John Pounsford, Frenchay Hospital, Bristol

Dr Mark Sculpher, University of York

Dr Iqbal Sram, NHS Executive, North West Region

Past members

Professor John Farndon, University of Bristol*

Professor Senga Bond, University of Newcastle-upon-Tyne

Professor Ian Cameron, Southeast Thames Regional Health Authority

Ms Lynne Clemence, Mid-Kent Health Care Trust

Professor Cam Donaldson, University of Aberdeen

Professor Richard Ellis, St James's University Hospital, Leeds

Mr Ian Hammond, Bedford & Shires Health & Care NHS Trust

Professor Adrian Harris, Churchill Hospital, Oxford

Dr Gwyneth Lewis, Department of Health

Mrs Wilma MacPherson, St Thomas's & Guy's Hospitals, London

Dr Chris McCall, General Practitioner, Dorset

Professor Alan McGregor, St Thomas's Hospital, London

Professor Jon Nicholl, University of Sheffield

Professor John Norman, University of Southampton

Professor Michael Sheppard, Queen Elizabeth Hospital, Birmingham

Professor Gordon Stirrat, St Michael's Hospital, Bristol

Dr William Tarnow-Mordi, University of Dundee

Professor Kenneth Taylor, Hammersmith Hospital, London

Diagnostics and Imaging Panel

Current members

Chair:
Professor Mike Smith, University of Leeds

Dr Philip J Ayres, Leeds Teaching Hospitals NHS Trust

Dr Paul Collinson, Mayday University Hospital, Thornton Heath

Dr Barry Cookson, Public Health Laboratory Service, Colindale

Professor David C Cumberland, University of Sheffield

Professor Adrian Dixon, University of Cambridge

Mr Steve Ebdon-Jackson, Department of Health

Mrs Maggie Fitchett, Association of Cytogeneticists, Oxford

Dr Peter Howlett, Portsmouth Hospitals NHS Trust

Professor Alistair McGuire, City University, London

Dr Andrew Moore, Editor, *Bandolier*

Dr Peter Moore, Science Writer, Ashtead

Professor Chris Price, London Hospital Medical School

Dr William Rosenberg, University of Southampton

Dr Gillian Vivian, Royal Cornwall Hospitals Trust

Dr Greg Warner, General Practitioner, Hampshire

Past members

Professor Michael Maisey, Guy's & St Thomas's Hospitals, London*

Professor Andrew Adam, Guy's, King's & St Thomas's School of Medicine & Dentistry, London

Dr Pat Cooke, RDRD, Trent Regional Health Authority

Ms Julia Davison, St Bartholomew's Hospital, London

Professor MA Ferguson-Smith, University of Cambridge

Dr Mansel Hacney, University of Manchester

Professor Sean Hilton, St George's Hospital Medical School, London

Mr John Hutton, MEDTAP International Inc., London

Professor Donald Jeffries, St Bartholomew's Hospital, London

Dr Ian Reynolds, Nottingham Health Authority

Professor Colin Roberts, University of Wales College of Medicine

Miss Annette Sergeant, Chase Farm Hospital, Enfield

Professor John Stuart, University of Birmingham

Dr Ala Szczepura, University of Warwick

Mr Stephen Thornton, Cambridge & Huntingdon Health Commission

Dr Jo Walsworth-Bell, South Staffordshire Health Authority

* Previous Chair

continued

continued

Methodology Panel

Current members

Chair:
Professor Martin Buxton, Brunel University

Professor Doug Altman, Institute of Health Sciences, Oxford

Dr David Armstrong, Guy's, King's & St Thomas's School of Medicine & Dentistry, London

Professor Nick Black, London School of Hygiene & Tropical Medicine

Professor Ann Bowling, University College London Medical School

Dr Mike Clarke, University of Oxford

Professor Michael Drummond, University of York

Dr Vikki Entwistle, University of Aberdeen

Professor Ewan Ferlie, Imperial College, London

Professor Ray Fitzpatrick, University of Oxford

Professor Jeremy Grimshaw, University of Aberdeen

Dr Stephen Harrison, University of Leeds

Mr John Henderson, Department of Health

Professor Richard Lilford, Regional Director, R&D, West Midlands

Professor Theresa Marteau, Guy's, King's & St Thomas's School of Medicine & Dentistry, London

Dr Henry McQuay, University of Oxford

Dr Nick Payne, University of Sheffield

Professor Margaret Pearson, NHS Executive North West

Professor David Sackett, Centre for Evidence Based Medicine, Oxford

Dr PAG Sandercock, University of Edinburgh

Dr David Spiegelhalter, Institute of Public Health, Cambridge

Professor Joy Townsend, University of Hertfordshire

Past members

Professor Anthony Culyer, University of York *

Professor Michael Baum, Royal Marsden Hospital

Dr Rory Collins, University of Oxford

Professor George Davey-Smith, University of Bristol

Professor Stephen Frankel, University of Bristol

Mr Philip Hewitson, Leeds FHSA

Mr Nick Mays, King's Fund, London

Professor Ian Russell, University of York

Dr Maurice Slevin, St Bartholomew's Hospital, London

Professor Charles Warlow, Western General Hospital, Edinburgh

Pharmaceutical Panel

Current members

Chair:
Professor Tom Walley, University of Liverpool

Dr Felicity Gabbay, Transcrip Ltd

Mr Peter Golightly, Leicester Royal Infirmary

Dr Alastair Gray, Health Economics Research Unit, University of Oxford

Professor Rod Griffiths, NHS Executive West Midlands

Mrs Jeanette Howe, Department of Health

Professor Trevor Jones, ABPI, London

Ms Sally Knight, Lister Hospital, Stevenage

Dr Andrew Mortimore, Southampton & SW Hants Health Authority

Mr Nigel Offen, Essex Rivers Healthcare, Colchester

Mrs Marianne Rigge, The College of Health, London

Mr Simon Robbins, Camden & Islington Health Authority, London

Dr Frances Rotblat, Medicines Control Agency

Dr Eamonn Sheridan, St James's University Hospital, Leeds

Mrs Katrina Simister, Liverpool Health Authority

Dr Ross Taylor, University of Aberdeen

Past members

Professor Michael Rawlins, University of Newcastle-upon-Tyne*

Dr Colin Bradley, University of Birmingham

Professor Alasdair Breckenridge, RDRD, Northwest Regional Health Authority

Ms Christine Clark, Hope Hospital, Salford

Mrs Julie Dent, Ealing, Hammersmith & Hounslow Health Authority, London

Mr Barrie Dowdeswell, Royal Victoria Infirmary, Newcastle-upon-Tyne

Dr Tim Elliott, Department of Health

Dr Desmond Fitzgerald, Mere, Bucklow Hill, Cheshire

Professor Keith Gull, University of Manchester

Dr Keith Jones, Medicines Control Agency

Dr John Posnett, University of York

Dr Tim van Zwanenberg, Northern Regional Health Authority

Dr Kent Woods, RDRD, Trent RO, Sheffield

* Previous Chair

Population Screening Panel

Current members

Chair:
Professor Sir John Grimley Evans,
Radcliffe Infirmary, Oxford

Ms Stella Burnside,
Altnagelvin Hospitals Trust, Londonderry

Mr John Cairns,
University of Aberdeen

Professor Howard Cuckle,
University of Leeds

Dr Carol Dezateux,
Institute of Child Health, London

Dr Anne Dixon Brown,
NHS Executive, Anglia & Oxford

Professor Dian Donnai,
St Mary's Hospital, Manchester

Dr Tom Fahey,
University of Bristol

Mrs Gillian Fletcher,
National Childbirth Trust

Dr JA Muir Gray,
Institute of Health Sciences, Oxford

Professor Alexander Markham,
St James's University Hospital, Leeds

Dr Ann McPherson,
General Practitioner, Oxford

Dr Susan Moss,
Institute of Cancer Research

Dr Sarah Stewart-Brown,
University of Oxford

Past members

Dr Sheila Adam,
Department of Health*

Professor George Freeman,
Charing Cross & Westminster Medical School, London

Dr Mike Gill, Brent & Harrow Health Authority

Dr Anne Ludbrook,
University of Aberdeen

Professor Theresa Marteau,
Guy's, King's & St Thomas's School of Medicine & Dentistry, London

Professor Catherine Peckham,
Institute of Child Health, London

Dr Connie Smith,
Parkside NHS Trust, London

Ms Polly Toynbee, Journalist

Professor Nick Wald,
University of London

Professor Ciaran Woodman,
Centre for Cancer Epidemiology, Manchester

Primary and Community Care Panel

Current members

Chair:
Dr John Tripp,
Royal Devon & Exeter Healthcare NHS Trust

Mr Kevin Barton,
East London & City Health Authority

Professor John Bond,
University of Newcastle-upon-Tyne

Dr John Brazier,
University of Sheffield

Ms Judith Brodie,
Age Concern, London

Mr Shaun Brogan,
Daventry & South Northants Primary Care Alliance

Mr Joe Corkill,
National Association for Patient Participation

Dr Nicky Cullum,
University of York

Professor Pam Enderby,
University of Sheffield

Mr Andrew Farmer,
Institute of Health Sciences, Oxford

Professor Richard Hobbs,
University of Birmingham

Professor Allen Hutchinson,
University of Sheffield

Dr Phillip Leech,
Department of Health

Dr Aidan Macfarlane,
Oxfordshire Health Authority

Professor David Mant,
Institute of Health Sciences, Oxford

Dr Chris McCall,
General Practitioner, Dorset

Dr Robert Peveler,
University of Southampton

Professor Jennie Popay,
University of Salford

Ms Hilary Scott,
Tower Hamlets Healthcare NHS Trust, London

Dr Ken Stein,
North & East Devon Health Authority

Past members

Professor Angela Coulter,
King's Fund, London*

Professor Martin Roland,
University of Manchester*

Dr Simon Allison,
University of Nottingham

Professor Shah Ebrahim,
Royal Free Hospital, London

Ms Cathy Gritzner,
King's Fund, London

Professor Andrew Haines,
RDRD, North Thames Regional Health Authority

Dr Nicholas Hicks,
Oxfordshire Health Authority

Mr Edward Jones,
Rochdale FHSA

Professor Roger Jones,
Guy's, King's & St Thomas's School of Medicine & Dentistry, London

Mr Lionel Joyce,
Chief Executive, Newcastle City Health NHS Trust

Professor Martin Knapp,
London School of Economics & Political Science

Professor Karen Luker,
University of Liverpool

Dr Fiona Moss,
Thames Postgraduate Medical & Dental Education

Professor Dianne Newham,
King's College London

Professor Gillian Parker,
University of Leicester

Dr Mary Renfrew,
University of Oxford

* Previous Chair

continued

National Coordinating Centre for Health Technology Assessment, Advisory Group

Current members

Chair:
Professor John Gabbay,
Wessex Institute for Health Research & Development

Professor Mike Drummond,
Centre for Health Economics, University of York

Ms Lynn Kerridge,
Wessex Institute for Health Research & Development

Dr Ruairidh Milne,
Wessex Institute for Health Research & Development

Ms Kay Pattison,
Research & Development Directorate, NHS Executive

Professor James Raftery,
Health Economics Unit, University of Birmingham

Professor Ian Russell,
Department of Health Sciences & Clinical Evaluation, University of York

Dr Ken Stein,
North & East Devon Health Authority

Professor Andrew Stevens,
Department of Public Health & Epidemiology, University of Birmingham

Past member

Dr Paul Roderick,
Wessex Institute for Health Research & Development

HTA Commissioning Board

Current members

Chair:
Professor Charles Florey,
Department of Epidemiology & Public Health, Ninewells Hospital & Medical School, University of Dundee

Professor Doug Altman,
Director of ICRF/NHS Centre for Statistics in Medicine, Oxford

Professor John Bond,
Professor of Health Services Research, University of Newcastle-upon-Tyne

Mr Peter Bower,
Independent Health Advisor, Newcastle-upon-Tyne

Ms Christine Clark,
Honorary Research Pharmacist, Hope Hospital, Salford

Professor Shah Ebrahim,
Professor of Epidemiology of Ageing, University of Bristol

Professor Martin Eccles,
Professor of Clinical Effectiveness, University of Newcastle-upon-Tyne

Dr Mike Gill,
Director of Public Health & Health Policy, Brent & Harrow Health Authority

Dr Alastair Gray,
Director, Health Economics Research Centre, University of Oxford

Professor Mark Haggard,
MRC Institute of Hearing Research

Dr Jenny Hewison,
Senior Lecturer, Department of Psychology, University of Leeds

Professor Sir Miles Irving
(Programme Director),
Professor of Surgery, University of Manchester, Hope Hospital, Salford

Professor Alison Kitson,
Director, Royal College of Nursing Institute

Dr Donna Lamping,
Senior Lecturer, Department of Public Health, London School of Hygiene & Tropical Medicine

Professor Alan Maynard,
Professor of Economics, University of York

Professor Jon Nicholl,
Director, Medical Care Research Unit, University of Sheffield

Professor Gillian Parker,
Nuffield Professor of Community Care, University of Leicester

Dr Tim Peters,
Reader in Medical Statistics, Department of Social Medicine, University of Bristol

Professor Martin Severs,
Professor in Elderly Health Care, Portsmouth University

Dr Sarah Stewart-Brown,
Director, Institute of Health Sciences, University of Oxford

Professor Ala Szczepura,
Director, Centre for Health Services Studies, University of Warwick

Dr Gillian Vivian,
Consultant, Royal Cornwall Hospitals Trust

Professor Graham Watt,
Department of General Practice, Woodside Health Centre, Glasgow

Professor Kent Woods,
Regional Director of R&D NHS Executive, Trent

Dr Jeremy Wyatt,
Senior Fellow, Health & Public Policy, School of Public Policy, University College, London

Past members

Professor Ian Russell,
Department of Health Sciences & Clinical Evaluation, University of York*

Professor David Cohen,
Professor of Health Economics, University of Glamorgan

Mr Barrie Dowdeswell,
Chief Executive, Royal Victoria Infirmary, Newcastle-upon-Tyne

Dr Michael Horlington,
Head of Corporate Licensing, Smith & Nephew Group Research Centre

Professor Martin Knapp,
Director, Personal Social Services Research Unit, London School of Economics & Political Science

Professor Theresa Marteau,
Director, Psychology & Genetics Research Group, Guy's, King's & St Thomas's School of Medicine & Dentistry, London

Professor Sally McIntyre,
MRC Medical Sociology Unit, Glasgow

Professor David Sackett,
Centre for Evidence Based Medicine, Oxford

Dr David Spiegelhalter,
MRC Biostatistics Unit, Institute of Public Health, Cambridge

Professor David Williams,
Department of Clinical Engineering, University of Liverpool

Dr Mark Williams,
Public Health Physician, Bristol

* Previous Chair

Health Technology Assessment 1999; Vol. 3: No. 5

Copies of this report can be obtained from:

The National Coordinating Centre for Health Technology Assessment,
Mailpoint 728, Boldrewood,
University of Southampton,
Southampton, SO16 7PX, UK.
Fax: +44 (0) 1703 595 639 Email: hta@soton.ac.uk
http://www.soton.ac.uk/~hta

ISSN 1366-5278