Statistical Methods in Medical Research 2010; **19**: 349–377

Trials in primary care: statistical issues in the design, conduct and evaluation of complex interventions*

GA Lancaster Postgraduate Statistics Centre, Department of Maths and Statistics, Fylde College, Lancaster, UK, **MJ Campbell** School of Health and Related Research, University of Sheffield, London, UK, **S Eldridge** Centre for Health Sciences, Queen Mary College, London, UK, **A Farrin** Clinical Trials Research Unit, University of Leeds, UK, **M Marchant** Public Health Department, East Sussex Downs & Weald Primary Care Trust, East Sussex, UK, **S Muller** School of Medicine, University of Manchester, UK, **R Perera** Department of Primary Health Care, University of Oxford, UK, **TJ Peters** Department of Community Based Medicine, University of Bristol, UK, **AT Prevost** Department of Public Health and Primary Care, Kings College, London, UK and **G Rait** Department of Primary Care and Population Health, University College, London, UK

Trials carried out in primary care typically involve complex interventions that require considerable planning if they are to be implemented successfully. The role of the statistician in promoting both robust study design and appropriate statistical analysis is an important contribution to a multi-disciplinary primary care research group. Issues in the design of complex interventions have been addressed in the Medical Research Council's new guidance document and over the past 7 years by the Royal Statistical Society's Primary Health Care Study Group. With the aim of raising the profile of statistics and building research capability in this area, particularly with respect to methodological issues, the study group meetings have covered a wide range of topics that have been of interest to statisticians and non-statisticians alike. The aim of this article is to provide an overview of the statistical issues that have arisen over the years related to the design and evaluation of trials in primary care, to provide useful examples and references for further study and ultimately to promote good practice in the conduct of complex interventions carried out in primary care and other health care settings. Throughout we have given particular emphasis to statistical issues related to the design of cluster randomised trials.

1 Introduction

Research in primary care has always presented particular challenges for researchers and the complexities of conducting such studies remain. Intervention studies in particular are often time consuming and require extensive planning and preparation. Complex interventions are conventionally made up of several interacting components

*Written on behalf of the RSS Primary Health Care Study Group.
After the first author's name, the names are in alphabetical order.
Address for correspondence: GA Lancaster, Postgraduate Statistics Centre, Department of Maths and Statistics, Fylde College, Lancaster LA1 4YF, UK. E-mail: g.lancaster@lancs.ac.uk

10.1177/0962280209359883

and present special problems related to their sensitivity to the local context and the logistics of applying experimental methods in a health care setting.[1] The Medical Research Council's (MRCs) new guidance document[1] on 'Developing and Evaluating Complex Interventions' outlines fully the processes to consider in developing, evaluating and implementing complex interventions. It provides a number of useful case studies for further reference, on issues such as theory-based reasoning, economic modelling, pilot and feasibility work, experimental design, reporting, user involvement and implementation. Campbell *et al*.[2] too emphasise the importance of thorough groundwork in designing and evaluating complex interventions and provide further case studies to highlight the importance of contextualising and conceptualising the problem at the development stage.

Cluster randomised trials in particular are common in primary care and are trials in which clusters (or groups) of individuals, rather than individuals themselves, are randomised. This requires active participation from general practices, nursing homes or households, for example, depending upon the unit of randomisation. The pressures of designing and successfully conducting primary care research, impact considerably on the role of the statistician in this context. With the recent developments of the UK Clinical Research Collaboration, the UK Clinical Research Network (UK CRN), the Primary Care Research Network (PCRN) and its local research networks, the provision to train, support and build capability in those working in this field is likely to be given increasing importance over the coming years. Parallel developments in the funding opportunities both for substantive applied research and methodological innovations of relevance to such research further emphasise the need for capacity building in this area.

In 2002, a group of statisticians set up the Primary Health Care Study Group under the auspices of the Royal Statistical Society, to concentrate specifically on issues related to primary care. The principal aims of the study group are to: bring together statisticians and researchers from a wide range of backgrounds who are working or have an interest in primary health care research; provide a forum for discussion of issues related to the design and analysis of complex interventions whilst learning more about the primary care environment in which the research takes place; improve the quality of quantitative research and facilitate collaborative working in this area. The latter aim is intended to be served by encouraging the development of a critical mass of expertise in both substantive and methodological areas of primary care research, resulting ultimately in raising the profile of statistics in primary care. The regular meetings of the study group have proved popular and effective in terms of disseminating ideas and encouraging networking, and it is as a consequence of these meetings that we decided to write this article to share the issues and ideas with a wider audience. The presentations from our meetings can be found on the Primstat data archive (at http://www.jiscmail.ac.uk).

The aim of this article is therefore to provide an overview of the statistical issues related to the design and evaluation of complex interventions, and to share useful examples and references for more in depth study of these issues, particularly in a primary care setting. Throughout where appropriate we have given particular emphasis to the statistical issues related to the design of cluster randomised trials. The intention is to build on existing guidance[1,2] in promoting good practice in the conduct of trials for assessing complex interventions in primary health care and other health care settings. The contribution of this article within the guidance framework is overviewed in Table 1.

Table 1 Overview of key statistical design issues in the development and evaluation of complex interventions in primary care and other health care settings

Phases given in MRC guidance framework[a]	Key elements in designing and evaluating complex interventions[b]	General points to consider	Key statistical design issues addressed in this article
Development	Background and context. (For more information and examples see MRC[1] and Campbell *et al.*[2])	Socio–economic background; Underlying cultural assumptions; Health service system; Government initiatives; Preventative policies.	
	Defining and understanding the problem. (See above[1,2])	Prevalence of condition; Population most affected; How condition is caused/ sustained; Potential for intervention and improvement.	
	Conceptualising the problem. (See above[1,2])	Levels of complexity of health problem and co-morbidity; Risk factors and factors influencing changes over time; Patient beliefs, symptoms and adherence to treatment.	
	Gathering evidence.	Systematic reviews; Epidemiological research; Qualitative research; Expert opinion.	Using evidence from primary studies, systematic reviews and qualitative studies to inform study design.
	Developing the intervention.	Identify key processes and mechanisms for delivery; Potential beneficial effect; Define target group; Optimise best treatment combinations.	Conducting primary care research in the UK: complying with research governance and assessing quality of care using the Quality and Outcomes Framework.
Evaluation	Developing and optimising trial parameters.	Testing the feasibility and and integrity of the trial protocol; Consideration of appropriate primary/ secondary endpoints; Recruitment and retention strategies;	Pilot studies and; pre-trial modelling; Selection of outcome measures for effectiveness and quality; Recruitment of practices and participants;

(*continued*)

Table 1 Continued

Phases given in MRC guidance framework[a]	Key elements in designing and evaluating complex interventions[b]	General points to consider	Key statistical design issues addressed in this article
		Method of randomisation to minimise imbalance;	Choosing the method of randomisation;
		Sample size considerations.	Sample size and between trial variation.
	Data collection and analysis.	Data collection forms;	
		Design of database;	
		Monitoring procedures;	
		Awareness of issues of data analysis for different study designs.	Choosing the method of analysis: cluster specific versus marginal models.
Implementation	Getting evidence into practice. (See new MRC guidance document[1])	Publication and dissemination strategy;	
		Stakeholder involvement;	
		Benefits, harms, costs for decision making;	
		Recommendations.	

[a] Phases of framework taken from the MRC, developing and evaluating complex interventions: new guidance document[1].
[b] Key elements adapted from Campbell *et al*. Designing and evaluating complex interventions to improve health care[2].

The complexities of conducting this kind of research inevitably require the careful planning and organisation of procedures well in advance of the launch of the main study. In this respect the design and analysis of previously conducted primary studies, systematic reviews and qualitative studies in the area of interest can greatly assist in planning the new study, and in ensuring its generalisability, and this is highlighted in Section 2. Generalisability will also be addressed by research governance bodies before permission for the study is granted. Compliance with research governance, the role of the statistician on ethics committees and particular intricacies of the primary health care system are addressed in Section 3. Testing the feasibility and integrity of the study protocol is another important aspect, and the ranges of concerns that may need pilot work are discussed in Section 4. One specific design issue is the selection of the most appropriate primary outcome measure; the related issues associated with different types of outcome measures are discussed in Section 5. Large-scale recruitment is common in primary care and needs attention to detail to be successful, as discussed in Section 6. Issues concerning randomisation of individuals or clusters of individuals are considered in Section 7. When clustering is involved then there is an additional component to the sample size calculation, called the intra-cluster correlation coefficient (ICC), and this is explained in Section 8. Statistical methods for planning and conducting the main analysis of the trial are outlined in Section 9, with a particular focus on the analysis of cluster randomised trials.

2 Using evidence from primary studies, systematic reviews and qualitative studies to inform study design

Medical research aims to improve the healthcare that is provided by practitioners and received by individuals. Unfortunately, much valuable research of high quality and good methodological rigour is wasted either because it lacks generalisability (external validity) or because it cannot be implemented in practice.[3]

2.1 Context and setting

The issue of generalisability provides a strong argument for primary studies of interventions to be carried out in the setting where they are most likely to be implemented, which is often in primary care, and is an important consideration when designing a trial. For example, current evidence of the effectiveness of interventions to lower blood pressure following stroke comes mainly from studies based in secondary care.[4] Therefore, if the results were to be used to plan a study in primary care, then there would be questions about the accuracy of the estimated effect size in this setting. However, information on the use of oral anticoagulants in atrial fibrillation is based on studies of the elderly and the majority of these interventions are carried out in a primary care setting.[5] In this case we can be more confident of using these results to plan similar future studies in a primary care setting. Similarly, an equally important issue is the need for adequate detailed descriptions of how the interventions being evaluated should be implemented in practice, in terms of who should do it, how it should be delivered and for what duration it is needed in the setting of interest. Descriptions tend to be vague and minimal, partly due to journal word constraints, but the recent option to have extended manuscripts appearing online should help to resolve this particular issue in time.

2.2 Systematic reviews

Systematic reviews based on clearly formulated research questions that identify relevant studies, appraise their quality, and summarise the evidence using explicit methodology are recognised as providing an essential underpinning for most empirical studies. Statistical approaches for systematic reviews of therapeutic interventions are well established.[6] Methods include meta-regression on the dose-effect relationship and Bayesian methods to identify adjusted trial specific effects, which have been shown to be useful tools in this context.[7,8] The feasibility and acceptability of an intervention can be measured by the level of attrition found in the study in those that are willing to participate. Information about low participation rates can also provide information about the feasibility and generalisability of uptake of the intervention. A systematic review allows the exploration of this across several studies, which are generally randomised controlled trials. 'Relative attrition'[9] (attrition in the intervention group divided by the attrition in the control group) has been used as a measure of acceptability that allows the calculation of an overall effect estimate and the study of different levels of attrition based on the population of each trial. Statistical methods for systematic reviews

of diagnostic test studies[10–12] and method comparison studies[13] are currently being developed, and provide ways of assessing agreement between two potential methods of measurement – for example type of thermometer to measure temperature in children,[14] in order to select the most appropriate technique to use in the main study. The same methodology used in the primary studies in these reviews may also be used to assess agreement between two assessors using the same measurement method on the same set of patients to ensure consistency in measurement, both concurrently and on different occasions.

Although the internal validity (i.e. the extent to which any observed difference between groups is due to the effect of the intervention being tested) of trials and systematic reviews has greatly improved, the findings have had a relatively small impact on guideline development. Consequently guidelines have had limited impact in aiding researchers in planning new studies on similar or related interventions. This is partly due to the fact that guideline developers want: (i) trials with consistent results, with large enough subgroups; (ii) individualised epidemiological data for risk based decisions, such as in the Framingham study for cardiovascular disease; (iii) research covering relevant populations such as ethnicities where health risks vary; and (iv) research in comparable settings, such as primary care.[15,16] Some advances in this area have been made with the creation of large databases for specific populations.[17] However, it is important to consider that guidelines generally have a national bias,[18] a median lag of 8 years in the research being used,[19] and a variable use of systematic reviews.[20]

2.3 Qualitative studies

Given the importance of the acceptability of interventions in health services research generally, and arguably particularly in primary care, quantitative methods alone are in general not sufficient to address all aspects of the evaluation of interventions in this context. Especially for complex interventions, a multiple integrated mixed methods approach which includes qualitative research methods is a key element and is of growing importance. Qualitative approaches can be used before a trial to explore issues related to the design of the trial, during a trial to unpack the processes of implementation and change and after the trial to explore reasons for the findings of the trial in line with the underlying theory.[21] Such methods are clearly not the primary expertise of those in the study group, but increasingly when the application of statistical methods in the design, conduct and reporting of such studies is being conducted in parallel with qualitative investigations, then both paradigms can influence the other at many stages of the investigations.[22] Specifically, qualitative methods have potential in the design of interventions (particularly complex interventions),[23] recruitment to trials,[24] understanding random allocation,[25] comparisons with patient reported outcomes,[26] assessing the acceptability of the outcome to both deliverers and receivers[27,28] and in the evaluation of the intervention itself – the latter both as assessments in their own right and also to play a role in interpreting the overall (qualitative and quantitative) findings from a trial.[29] Various techniques are available, and qualitative studies are valuable both for the participants and health professionals involved in the study.[22]

3 Conducting primary care research in the UK: complying with research governance and assessing quality of care using the Quality and Outcomes Framework

In general the public perceives academic research to be honest and credible, and every effort is made within the research community to maintain this reputation. Governance in research is about ensuring research integrity in order that the public maintains confidence in the processes used and the resulting outcomes, that research participants are protected from abuse and researchers are protected from accusations of misconduct.[30] In order to achieve high quality clinical research and patient care, researchers need to comply with a wide range of legal requirements, including the European Clinical Trials Directive and the Data Protection Act (see www.ukcrn.org.uk/index/clinical/regulatory.html). The second edition of the National Health Service (NHS) Research Government Framework[31] was published in 2005 and covers all aspects of research governance, including the responsibilities of the investigators, funders and sponsors, standards, monitoring and inspection and research ethics, with the aim of promoting a quality research culture.

Primary health care within the UK differs markedly in practice from most other countries, and in order to meet the challenges of conducting primary care research in the UK, we need to understand the primary health care system, and the way that quality of care is monitored through the Quality and Outcomes Framework (QOF).

3.1 Conducting primary care research in the UK

General Practitioners (GPs) in the UK are usually self-employed or work within a limited company, and they contract to the NHS to provide specific services to a given number of patients. There are few GPs who now work single-handedly, the move over recent years being towards primary health care centres incorporating partnerships of GPs as well as other health care staff – for example, practice nurses, health visitors, counsellors and chiropodists. This infrastructure presents particular types of problems in the design and analysis of research projects as there will virtually always be a need to include a number of GP centres. A further complication arises from the way the NHS is structured across the UK. Primary Care Trusts (PCTs) are the NHS bodies that commission GPs' services within their areas, which are normally coterminous with counties or unitary authorities. They are also responsible for facilitating research locally within the NHS and for ensuring the integrity of the research and its practitioners. Problems have arisen in setting up research projects across different areas as each of the PCTs involved needs to be assured by the researchers of their ability to conduct their projects appropriately. Researchers need to take time to present to PCTs the aims and plans of the proposed research in order to obtain their approval, which may take well up to 6 months, and occasionally even longer.[30]

The National Institute for Health Research (NIHR) has recently published guidance in their 'Research in the NHS – Human Resources Good Practice Resource Pack'[32] to help both the PCTs and Higher Education Institutions (HEIs) to be aware of their responsibilities. Honorary contracts are required for researchers where their work concerns NHS patients, and Criminal Records Bureau checks will be needed particularly

when children or vulnerable adults are to be involved.[33] It is, however, sufficient to have such a contract with only one NHS organisation even though the research may be conducted elsewhere in the NHS. A Research Passport system[32] has recently been introduced to streamline the application process and is needed even when the research is not concerned with patients' care. Undergraduate and postgraduate students proposing to undertake research need to be supervised by HEI staff with honorary research contracts. HEI-employed GPs, practising within the NHS, also need such contracts if the research extends beyond their normal clinical practice. HEIs, should ensure that research is adequately planned, and researchers adequately resourced and funded. Indemnity from harm due to clinical negligence in research is, however, the responsibility of the NHS once the honorary contract is awarded.[32]

Any research involving NHS patients or healthy volunteers, their patient records or human tissue, also needs approval from a research ethics committee (REC). The NHS Patient Safety Agency is responsible for the research ethics service[31] (see www.nres.npsa.nhs.uk) through which application for ethical approval is made. RECs are responsible for acting primarily in the interest of potential research participants, protecting their rights, dignity, well-being and safety. Researchers, for instance, have no right of access to an identifiable individual patient's record without explicit consent from the individual. It is RECs that permit researchers to approach individuals for such consent – see for example, the recent Wellcome guidance on 'The use of patient data from GP records for research'.[34] The safety and interests of researchers and the appropriateness of the proposed research is also taken into consideration by RECs before granting approval. Furthermore, RECs do examine statistical issues, the sample size being proposed as well as the planned analysis. A summary of any advice given to researchers by statisticians is often requested by RECs at the time of application, and RECs do sometimes include statisticians among their membership.[35] There is, however, debate about whether RECs should examine statistical issues because although in some institutions scientific review does take place prior to submission of the proposal to REC, this is not always the case and the two processes are often run in parallel.

3.2 The research potential of Quality and Outcomes Framework indicators

Before the new contract was agreed with GPs in 2004 the NHS was very restricted in its ability to monitor the quality of care provided to patients. This was particularly of concern in the more deprived areas of the UK.[36] Financial incentives relating to up to 30% of GP income, have been introduced through the achievement of targets monitored in the QOF. There are five domains: clinical care; organisation; patient experience; education and training; and other additional services provided that are covered by the indicators. Clinical care covers 19 areas from coronary disease and cancer to obesity and smoking habits. Points are awarded according to estimated workload needed to achieve targets and prevalence of disease (dependant on age, sex and deprivation level). The points are then converted into payment, adjusting for practices' list sizes and comparison with national averages. Very few GPs are currently not covered by the QOF but there have been modifications and additions since the first set of indicators was created for 2004/5 (see http://www.qof.ic.nhs.uk for further details).

When using the QOF to assess differences in quality of care one should be aware of certain issues that may affect comparisons between practices. Patients may be excluded for various reasons from the assessments – for example, because of patients' failure to attend for reviews, frailty of condition or refusal of treatment. GPs may differ in how certain patients' conditions are recorded or how interventions are assessed, such as advice to quit smoking and its outcome. Practices may also change in composition over time with new staff introducing new skills, which can affect the quality of care provided. Similar questions arise about equity of care (for example, across different areas and socio–demographic groups), and these may also change over time. Finally, it has to be remembered that the QOF is primarily a payments-driven system and was not therefore created specifically for research purposes.[36]

4 Pilot studies and pre-trial modelling

Planning a trial is time-consuming and costly but due to the nature and enormity of many of the trials in primary care it is an extremely important pre-requisite of most major funding bodies. Lancaster *et al.*[37] showed that there was a dearth of pilot studies in the literature that stated they were specifically in preparation for a randomised controlled trial, and that give a clear list of key objectives relating to the pilot phase. Often feasibility studies are conducted as small stand-alone studies on an ad-hoc basis, and in addition they may be subject to publication bias and therefore not readily available as learning tools for researchers. The problem is compounded when we are dealing with large clusters of people, as these studies require a great deal of administrative effort and co-ordination, often including marketing and public liaison work to promote the importance of the study to key stakeholders. This type of preparation and attention to detail can improve recruitment (see Case study 14 in MRC guidance document)[1]. Funding for pilot studies for trials is also an issue. It is difficult to get funding for a pilot study, although sometimes local NHS Trusts do allocate Research and Development funds in this respect. In addition, large funders such as the Health Technology Assessment do allow suitable costs for a pilot/feasibility phase before the main trial is conducted.

Pilot studies are conducted to assess the feasibility and integrity of the study protocol in a variety of ways[37] as shown in Table 2. One important objective of a pilot study is to test the acceptability and feasibility of developing and implementing the intervention,[1] illustrated by the PANDA trial[38,39] case study in Table 3. In this respect pre-trial modelling has been shown to be very useful in assessing the likelihood of the intervention being effective. Eldridge *et al.*[40] developed a cost-effectiveness model of a complex intervention from pilot study data in order to inform the viability and design of a subsequent falls prevention trial. Using two models, they first estimated the probability of falling over a 12-month period based on a probability tree, and then used Markov simulation to assess the impact of the programme over time. The first model indicated that the intervention would reduce the proportion falling by only 2.8% over a 12-month period. The major reason for this small effect was that less than a quarter of older people at risk of falling were assessed using their screening tool. Sensitivity analyses showed that the only scenarios that produced a substantial increase in the

Table 2 Key objectives of a pilot study

Objective	Issues to consider
Test the integrity of the study protocol.	Important if multiple sites are to be involved; Inclusion/exclusion criteria; Are interim analyses necessary? Would a pilot RCT be helpful?
Sample size calculation.	Obtain initial estimates, considering variability; Estimate of intraclass correlation coefficient if to be cluster randomised.
Recruitment and consent rates.	Important for planning the length of the study; Strategy for recruitment of practices and participants; How to best explain the study in layman's terms; Readability of patient information and consent forms.
Develop and test the implementation and delivery of the intervention.	Drug preparation, storage; Duration of delivery; Is it possible to adhere to a dietary regime? Testing of materials, equipment and techniques; Is self-administration possible? Is an on-call help service needed?
Determine the acceptability of the intervention.	To participants, assessors and funders; Side effects; Feasibility of costs; Pre-trial modelling of cost-effectiveness.
Train staff in delivery and assessment procedures.	Inter-rater and intra-rater reliability; Calibration of instruments; Data collection, recording and data entry.
Selection of most appropriate primary outcome measure (endpoint).	Use more than one primary outcome measure? Secondary outcome measures; Use of biomarkers? Patient-reported outcomes?
Randomisation procedure.	How to implement; Is use of 24 h randomisation service needed? Acceptability to participants.
Pilot data collection forms and/or questionnaires.	Face/content validity; Self-completion at hospital or at home; Use of postal questionnaires; Are home visits and interviewers required?
Prepare and plan data collection and monitoring procedures.	Database – data entry, validation methods, backup; Forms for monitoring adverse events, missing data.

effect of the intervention were those in which all older people are assessed, and this was not cost-effective. They found that even if policy-makers were willing to spend 30 000 pounds per quality-adjusted life-year gained, there was only a 40% chance that the intervention would be cost-effective.

A second objective of pilot studies is the selection of the most appropriate outcome measure for assessing the effect of the intervention. Here we want to consider several suitable outcomes, however, a new measure should only be used as a primary outcome

Table 3 Case study 1

Case study 1 – Developing and testing the acceptability of the intervention for the PANDA (Patients and Decision Aids) trial[38]

Objective: To devise and test a Patient Decision Aid (PDA) to help patients with Type 2 diabetes decide when they should go on to insulin.

The premise was that patients who made fully informed decisions were more likely to adhere to treatment and be better controlled.

Method: Feasibility study to develop the PANDA study protocol following the MRC complex intervention guidelines and comprising:

(i) Expert Panel of GPs, nurses, a diabetologist, patients, statistician, clinical decision experts;
(ii) Needs assessment exercise involving a Review Panel of 14 GPs and nurses and nine patients with Type 2 diabetes;
(iii) Systematic review of the evidence for insulin therapy;
(iv) Drafting of the decision support intervention using the Ottawa Guidelines;[39]
(v) Review of the intervention by the Review Panel;
(vi) Pilot study to test the decision aid for acceptability and feasibility;
(vii) Sample size calculation: 30 practices (15 to receive the PDA and 15 to receive usual care) and with 15 patients per practice to have 80% power to detect a difference of 0.5% HBA1c at the 5% significance level.

Results: The feasibility study revealed that general practitioners found the PDA acceptable, and that there were sufficient eligible and willing patients in the practices to devise a cluster trial of the intervention against usual care in one city using 30 practices.

if it has been shown to be valid and reliable in the population in which it is to be used before its use in the main study, and a pilot study offers an opportunity to do this. This is particularly important when children are involved as often outcome measures have been tested for use in adults but not necessarily for use in children.[41] Another objective of a pilot study is to obtain information and initial estimates to use in sample size calculations.[37] In studies involving clustering, when there is no information on the primary outcome, then it may be necessary to consider a range of different scenarios to see how each would affect the power of the study, for example, in a sensitivity analysis.[42] This should also include consideration of the recruitment rate and possible drop-outs and withdrawals. Piloting of data collection and follow-up forms is also recommended, particularly when there are multiple assessors or self-completion is required by the participants.

4.1 Example: UK BEAM trial

The value of well designed pilot or feasibility studies, prior to large multi-centre randomised controlled trials, is perfectly illustrated by the feasibility study for the UK Back pain, Exercise, Active management and Manipulation (UK BEAM) trial.[43] As seen in Table 4 this was planned as a primary care cluster randomised $3 \times 2 \times 2$ factorial trial[44,45] of three treatments for back pain, namely 'active management' randomised at the practice level, and 'spinal manipulation' and 'exercise classes' both randomised at the patient level. The preceding feasibility study was designed as one large study with a

Table 4 Case study 2

Case study 2 - Testing the integrity and feasibility of the trial protocol for the UK BEAM (Back pain, Exercise, Active management and Manipulation) trial
Objective: To test the integrity and feasibility of the trial protocol. The trial was planned to be a cluster randomised 3 × 2 × 2 factorial trial of three treatments for back pain ('active management' randomised at the practice level, and 'spinal manipulation' and 'exercise classes' randomised at the patient level).
Method: Particular issues investigated concerned: (i) Sample size – primary outcome variability and ICC estimates, and identification, recruitment and follow-up rates; (ii) Data collection processes – piloting of forms, length of patient questionnaire, most effective method to collect practice data; randomised sub-study of patient diaries versus questionnaires to compare methods for collecting economic data; (iii) Staff training – acceptability of multi-disciplinary approach, effectiveness of generic primary care educator versus clinical back pain expert; (iv) Treatment delivery – sufficient number of therapists, and manipulators and locations for treatment;
Results: The majority of the methods piloted were feasible. Differential recruitment between clusters led to the abandonment of the planned randomisation at practice level. Intervention practices recruited more than twice as many participants as control practices and participants in intervention practices differed significantly at baseline to those in control practices, having, for example, milder back pain, less physical limitation, less depression, higher percentage in full-time work and higher levels of educational attainment. The protocol was changed so that all patients were individually randomised and all practices were trained in active management to maximise recruitment.

number of planned embedded sub-studies. The aim was to pilot all aspects of the trial including the intervention,[23] to identify problems in design or execution,[46] investigate unresolved issues and demonstrate that the main trial could fulfil its aims, in terms of its design and implementation, and was therefore worthy of funding.

Overall, the feasibility study demonstrated that the majority of methods and processes were successful. It identified where changes were required to the trial design or execution and highlighted unexpected problems,[46] allowing further design changes before the start of the main trial. This study proved that pilot studies are vital, especially when evaluating complex interventions,[1] for providing planning information and identifying unanticipated issues in advance of expensive, complex trials. This is of course not to say that all pertinent issues will be resolved by such studies, as almost inevitably further issues will arise when the main study begins to recruit on a wider basis. A major factor is the additional workload for researchers and practice staff,[23] especially if the study involves sites some distance from the trial centre. Nonetheless, feasibility studies are crucial in keeping such difficulties to a minimum.

5 Selection of outcome measures for effectiveness and quality

In general outcome measures need to be valid (shown to have face/concurrent/predictive validity), repeatable (stable over time when disease state is not changing), reproducible

(when applied by different assessors), and objectively measured in situations where self-reporting is unreliable (for example, self-reported smoking cessation with additional biochemical validation).[41] The primary outcome of a trial may be directly measurable (for example, as the peak flow rate, blood pressure or severity of effusion), or it could be a patient-reported outcome,[47] where the person receiving the treatment describes how well they feel the treatment has worked. Patient-reported outcomes are commonly assessed using a questionnaire, where a series of closed questions, with dichotomous or categorical response options, are asked. The scores assigned to these categories are then generally summed to get a total score. A discussion of the analysis of such studies has been given by Lall *et al*.[48] One might also consider individually set outcome measures, whereby the patient suggests activities in which they wish to see improvements in their performance.[49]

Outcomes can be measured at the individual level, for example improvement in a patient's pain score, or at the level of a group or cluster, which in primary care research is usually the GP practice. In primary care an intervention administered at the practice level might be the introduction of a universal parenting programme to prevent early childhood behavioural problems, with main outcomes concerning parenting, child behaviour and maternal mental health.[50] In choosing the primary and secondary outcomes to be measured, it is imperative that they are suitable for the purpose, both theoretically and practically. Researchers must consider the most important outcome for a trial, and in primary care research there are likely to be many to choose from. For example, in a trial for knee pain,[51] it is likely that as well as a measurement of pain, knee function, ability to work and satisfaction with treatment may also be assessed. However, only one or two of these would be selected as the primary outcome measure(s) by which to assess the overall success of the intervention, and the required sample size calculation would be based on only the primary outcome measure(s).

Before deciding on an outcome measure, it is important to think about the type of statistical analyses required. Many physiological measurements will be interval in nature; that is, they will be measured on a linear scale where a change in score from 5 to 10 units represents the same difference as a change from 15 to 20 units – for example when measuring blood pressure. However, this is unlikely to be the case with patient reported outcomes such as Health Related Quality of Life (HRQL) because of the way the scores are created from the individual items. These scores often give the impression of having a linearly increasing scale; when, in fact, the scale is ordinal and so the arithmetic operations needed to find differences or means may not make sense.[52,53] Although interval-level measures are now beginning to appear in health research,[54,55] it is often the case that ordinal scores are used as trial outcomes and that arithmetic operations are carried out. It is possible to achieve a linear score from questionnaire data when the score is created using the Rasch measurement model.[56] Empirical evidence comparing the adoption of parametric with non-parametric approaches for ordinal data analysis with respect to HRQL, did find discrepancies in results between approaches.[57] Walters *et al*.[58] suggested that if the HRQL measure has a large number of ordered categories, most of which are occupied, and the underlying scale really is continuous but measured imperfectly by an instrument with a limited number of discrete values, then an informal rule of thumb is that this discrete scale should be treated as continuous if it has seven or more categories and as ordinal otherwise. More generally,

the use of HRQL measures is gaining increasing importance in the economic evaluation of interventions. Whenever possible a good approach is to use two HRQL measures: one that is disease-specific, such as the Hospital and Anxiety Depression Scale[59] for depression, and one that is generic, such as the Short Form 36,[60] (SF36). Disease specific HRQL measures sometimes must be developed in a pilot stage but it is wise to include popular generic scales as well, so that comparisons can be made across studies.

6 Recruitment of practices and participants

Recruitment to primary care studies is complex and response rates can be low. There are different levels of recruitment, including the recruitment of practices and recruitment of patients or participants. These often require tailored approaches, and may be trial or study specific. Eldridge *et al.*[61] give examples of different recruitment strategies and discuss what can be done to avoid bias in identifying and recruiting participants to cluster randomised trials, where recruitment may operate at a number of different levels. Ross *et al.*[62] identified barriers to patient participation that included concerns about information and consent; patient preferences for treatment and additional demands such as additional procedures. Barriers to clinician participation included lack of staff and training; concern about the impact on the doctor patient relationship; time constraints and difficulties with the consent procedure. Mapstone *et al.*[63] and Watson and Torgerson[64] carried out systematic reviews of strategies to improve recruitment and in both cases concluded that some strategies were not necessarily generalisable and as such it was not possible to predict their effect in different settings. They concluded that funders and researchers should ensure that evaluation of recruitment strategies is incorporated into research studies. The Bells Palsy trial[65] is a good example of devising strategies to maximise recruitment and retention. These included pilot work; the use of a good research network infrastructure in Scotland for sources for referral; and the involvement of a celebrity to promote the study in the media.[66]

Successful recruitment requires a co-ordinated and multi-faceted approach and, as we have already seen in Section 4, the piloting of trials in primary care is an important first step.[37,46] Pilot studies often address barriers to participation of practices, efficient ways to identify the sample within practices and to recruit participants, consent rates and whether completion of outcomes measures is acceptable to participants. It is important to engage practices in the study at an early stage. A research question that is considered to be important to primary care will usually make recruitment easier, as it will engender greater practice interest. Likewise it is important to consider how the study will impact (or at least be perceived to impact) on the patient–doctor relationship more generally, and the priority given to the research question in relation to other issues.[24] Providing training for practice staff either in an area of disease management or research will also be appreciated, especially in areas where patients are considered vulnerable, and some GPs may lack confidence in raising research issues within a sensitive consultation such as those relating to mental health.[24] Most importantly, recruitment of practices to national studies can be facilitated by primary care research networks and by identifying a named contact in the practice. These networks have existed for a number of years

at both national (see Medical Research Council General Practice Research Framework (MRC GPRF) found at http://www.gprf.mrc.ac.uk/) and local levels. More recently the UK CRN through the NIHR Clinical Research Network Co-ordinating Centre (NIHR CRN CC) for England has set up the Primary Care Research Network (PCRN) in England (http://www.ukcrn.org.uk/index/networks/primarycare.html) with eight local primary care networks being nationally co-ordinated. There are equivalent research networks in Scotland, Wales and Northern Ireland. Both the PCRN and MRC GPRF provide a recruitment framework and support for both local and national peer reviewed and funded studies. In a new incentive introduced by the Royal College of General Practitioners in conjunction with the NIHR CRN and PCRNs, practices involved in research can now seek accreditation through the 'Research Ready' accreditation scheme found at http://www.rcgp.org.uk/researchready.

Another important area for consideration is obtaining informed consent and there is a growing commitment to increase public awareness, understanding and support for the use of patient records in research. Whilst individual patient consent is not needed when records remain anonymised and unidentifiable, when records are identificable then consent is necessary and it is important to remember that patients can opt-out of participation in research (see Wellcome guidance document[34]). Response rates may be higher when patients are invited by their GPs to participate in a study, however, successful recruitment often means using people other than GPs to identify and consent possible participants. In fact there is evidence that studies where GPs are asked to gain consent are less likely to recruit.[67] Staff such as practice nurses or non-clinical practice staff can be trained to take informed consent and increasingly research network staff are becoming trained in the internationally recognised principles of Good Clinical Practice[68] (see http://www.mhra.gov.uk/Howweregulate/Medicines/Inspectionandstandards/GoodClinicalPractice/index.htm). A nursing model is used extensively in the MRC GPRF and is being considered by the PCRN. One way of minimising work for practice staff is to use electronic queries to identify samples and this approach is being developed within the ePCRN clinical trial management system (see www.ePCRN.org).[69] Other methods include using electronic systems to trigger an alert within a consultation[70] or templates on practice systems when eligible patients are seen. There is also evidence that having a strong 'brand' for the trial, having well-developed marketing strategies and engaging with the stakeholders (such as practices or participants) can help with recruitment. Practices need to be adequately reimbursed for the time taken to take part in research, and support for this is now available from the Comprehensive Local Research Networks.

7 Choosing the method of randomisation

The choice of a randomisation method usually requires greater consideration in the primary care setting. Firstly, there is greater use of cluster randomised designs in which either relatively few units, usually general practices[71] are randomised, or clusters, such as households are small.[72,73] With this type of design, there is a higher probability that arms in a study will become imbalanced in size, and that baseline covariates will show an imbalanced distribution at the level below the unit of randomisation, usually

individuals – or even, in studies with relatively few clusters, at the level of randomisation itself. Secondly, complex, multi-component, intervention trials are more commonly encountered in primary than secondary care, and in particular screening[73] and preventive trials[72] are more common. These complex trials often have multiple parts and multiple routes through which an intervention can act and operate. Consequently there are an increasing number of covariates that may predict outcome, and which need to be considered to avoid confounding. Thirdly, early phase explanatory trials may involve testing a greater number of types of factors such as multiple interventions, or different doses or levels of contact from an intervention deliverer. These may require for example factorial,[44,74] or other multi-arm trials[72,75] where covariate-imbalance is more probable over the large number of pairwise arm comparisons.

7.1 Imbalances in size of treatment groups

There are two issues to consider for cluster randomised trials to optimise power: ensuring equal numbers of clusters and ensuring equal numbers of patients per treatment arm. In studies where few units of observation are randomised there may be a concern that imbalances will lead to a reduction in study power. The likelihood of such an occurrence can be evaluated. For example, consider a two-arm parallel groups cluster randomised trial of 30 general practices, which assumes an equal allocation ratio for 80% power. By binomial sampling, there is a 10% chance of obtaining a 2 to 1 or more uneven allocation, which reduces the power below 75%. Blocking, the method of generating randomly permuted blocks to form the randomisation sequence, can be used to achieve the planned allocation ratio. Here interventions are assigned randomly within blocks to ensure balance within the blocks. Blocks of larger or varied sizes can be used to form a sequence that lowers predictability of successive assignments. Predictability is an important issue,[76,77] particularly with respect to allocation concealment to the participants. In many cluster trials, randomisation is done at the start and the trials are open such that everyone recruited in the same cluster gets the same allocation, which means that participants are only blind to allocation status as long as the actual treatment can remain concealed, for example by using a placebo, and this is usually not feasible (if it were, cluster randomisation would, in most cases, not be necessary). In individually randomised trials, the choice and concealment of block size and factors specific to the practical running of the trial, such as the arrival of batches of participants to be randomised in a varied order, may then alleviate concerns of predictability.

7.2 Imbalances in covariates

As already mentioned another consideration in primary care trials is the choice of a method that will best avoid imbalances in baseline covariates. At the design stage, a number of covariates may be thought to be important predictors of outcomes, and if the baseline distribution of these were to become imbalanced between study arms then this may affect the face validity of comparisons and conclusions. An analysis-stage adjustment for those covariates that show imbalance would allow correction according to an assumed model, but then the unadjusted and adjusted estimates of intervention effects may differ in size and statistical significance. This can be problematic

when interpreting the results, especially as a pre-specified statistical analysis plan with a defined primary model and covariates requires prior knowledge of the covariates that may become imbalanced.

A design-stage alternative is to eliminate the occurrence of imbalance in chosen covariates by using a restricted randomisation method such as stratification by covariates when there are sufficient randomisation units.[74] This uses blocking within the strata formed from all possible combinations, or interactions, of the categorical, or categorised continuous, covariates. Alternative methods include minimisation,[76] which balances the main effects of more variables than is feasible using stratification.[75] The idea here is that the next participant to enter the trial is allocated to the intervention group that would minimise the overall imbalance between the two groups according to a pre-specified set of covariates. The selection of one covariate as a stratifier immediately reduces the chance of extreme imbalances in another correlated covariate.

7.3 Practical considerations

In the primary care setting, care is also needed to ensure that randomisation is not undermined when, for example, a randomisation program or schedule has not been checked, sequential allocations are left unmonitored, or numbered sealed envelopes are used without sufficient consideration of the randomisation setting. In primary care, the use of envelopes has been common – for example, in the situation when previously unidentified eligible patients arrive in primary care for consent and randomisation. The NIHR CRN CC recommends the use of centralised randomisation services operating, for example, from within recognised Clinical Trials Units. Web-based solutions are also becoming available.[78] The CONSORT statements for individual and cluster randomised trials are informative for planning randomisation, covering issues such as procedure, method and level of randomisation.[79] The decision to randomise clusters requires justification because of the associated cost of the increased numbers of patients involved.[80] The avoidance of potential contamination between individuals in different study arms is a common justification.[81] The new NIHR network of regional Research Design Services can give advice regarding issues such as randomisation (see www.nihr.ac.uk/infrastructure).

The ProActive trial is an example of a complex intervention to increase physical activity in participants with parental history of Type 2 diabetes using three trial arms.[72] Covariate imbalance was an issue in this multi-arm trial such that the greater likelihood of pairwise imbalances in six identified covariates led to the decision to balance main effects using minimisation, importantly having a stochastic element for a more robust statistical basis and reduced predictability of assignment. In addition, two covariates were considered important enough to retain in interaction form via a single combined stratifier, and stratifers were weighted relatively in the method to improve likely effective balance, with weights informed using over-resampling of pilot sample participants' covariates.

Alternative design methods used in community based studies are matched-pairs,[82] or randomisation units matched in sets larger than size two,[83] and use of balanced randomisation with permutation or randomisation tests.[71,84] Designs that are efficient

in addressing multiple unrelated conditions require a carefully considered randomisation approach.[85] Altogether different approaches may be required for interventions that involve several intervention deliverers, depending on whether this person is common to all arms, or restricted to a subset.[72] Practical considerations can restrict the use of a further random allocation of a participant to an intervention deliverer. Sometimes consideration needs to be given as to whether randomisation will be possible, and in some cases patient preference designs may need to be examined.[86]

8 Sample size and between practice variation

The existence of variations between general practices in treatment and referral activity is well established.[87–89] This variation is relevant in many trials in primary care. For example, some studies will have an explicit aim to reduce variation or will at least have been conceived because of the existence of variation. When a trial is conducted in a single practice, or a small group of geographically close practices, between practice variation may have implications for the generalisability of the study. However, perhaps the most important implication of between practice variation for trials in primary care is in the design of cluster randomised trials,[90] since between cluster variation must be allowed for in sample size calculations of these studies. In primary care the most common type of cluster is the general practice – hence the importance of knowing about between practice variation in this context.

Understanding and measuring the extent of between practice variation for interpreting or planning studies depends on good quality, relevant data. There are a number of large data sets currently available that provide information on variability, for example, the General Practice Research Database (GPRD), Mediplus and MIQUEST. In addition, since the establishment of the RSS Primary Health Care Study Group in 2002, we have seen the development of Primary Care Trusts, and the introduction of the QOF (see Section 3.2), both of which have had the potential to improve the amount and quality of data, although the target setting associated with the QOF may also distort data.[36,91,92] Information on between practice variation can, of course, also be collected in primary research studies. Whilst estimates of variation taken from primary studies might be more closely related to the planned outcome of interest, they may be less precisely estimated when compared to estimates taken from large general databases where greater precision can be achieved. There is, therefore, a trade off to be made between bias and precision.

In addition to having robust data, investigators must also have a way of measuring between practice variation and make a decision about whether several measures can be combined. Sometimes suitable measures can be extracted from the QOF. If it is felt necessary to combine outcome measures, a decision must be made about how these are to be combined, for example by counting the number of positive scores for a patient. Different ways of combining measures can lead to different assessments of the extent of variation and the position of an individual practice within any quality ranking.[93] In relation to trials, as highlighted in Section 5, a decision must be taken about the primary outcome to be assessed. Between practice variation in the outcome will be relevant to the extent that this affects the sample size required. For a cluster randomised

trial, the measure of between practice variation usually adopted is the ICC and this can be used more generally to describe between practice variation.[94,95] The best way of thinking about an ICC is as the proportion of total variation between individuals that is explained by between practice variation (as opposed to variation within practices). Thus it ranges in value between 0 and 1. The actual value depends on a number of factors such as the type of cluster, the type of outcome, the setting and for binary outcomes, the overall outcome prevalence. For trials randomising general practices and involving patient-level outcomes, ICCs usually take values around 0.05. In recent years, a number of papers have been published which provide lists of ICCs relevant to primary care.[96–100]

Before conducting a randomised controlled trial, investigators should not only think about how to predict ICCs but whether a cluster randomised trial is the best design. These trials are often conducted to avoid contamination,[101] or because an individually randomised trial is impossible or to reflect the manner in which an intervention would ultimately be delivered. The sample size is commonly calculated by finding the number required by an individually randomised trial with the same effect size, significance level and power and multiplying it by the Inflation Factor or Design Effect. These are given by $1 + (n - 1)\rho$, where ρ is the ICC and n is the cluster size when cluster sizes are all equal. Since ρ is positive this has the effect of increasing the sample size compared to a comparable individually randomised study. When cluster sizes vary the mean cluster size can be used in this formula instead of n but it underestimates the sample size required.[102] An alternative formula which accounts for variable cluster size is conservative,[103] but shows that when the coefficient of variation of cluster size is less than 0.23 any effect on power is essentially negligible. Unfortunately, coefficients of variation of cluster size are usually much larger than this when general practices are randomised, so investigators should take account of variable cluster size when calculating sample size. Kerry and Bland[104] show the implications of variable cluster size on power for a range of scenarios, and also highlight the risk of having empty clusters, a particular problem in primary care where recruitment can be a problem. When studies recruit smaller cluster sizes than intended, investigators may include more clusters to compensate.

Unfortunately, many cluster randomised trials conducted before 2000 did not account for clustering correctly and were consequently underpowered.[105] The situation has improved since then. A number of recent publications make recommendations for improving the design, analysis and conduct of these trials. These include piloting to iron out execution issues,[46] using Bayesian methods to improve prior estimates of ICCs,[106,107] ensuring that bias does not occur at the stage when patients are identified and recruited,[108,109] and paying more attention to the generalisability of these trials.[105,108] In particular, Turner *et al.*[110] showed that an estimated ICC from a previous trial with the same target participant group (e.g. diabetes patients) and cluster (e.g. practice) typically has sizeable sampling variability, and there is therefore a question as to whether such an estimate will accurately reflect the ICC that will be observed, which itself has sizeable sampling variability. They used Bayesian and classical methods to form a distribution using the previous estimate and its standard error (or several of these, when available, via meta-analytic methods) in order to obtain, through simulation from this distribution, a sample of ICC values that were then each combined with a design option (of numbers of clusters and participants) which provided a distribution

for power. For a fixed average power (e.g. of 80%), they recommended designing trials that opt for larger numbers of smaller clusters rather than a smaller number of larger clusters, in order to reduce the impact of the sizeable sampling variability and thereby to lower the risk that the occurrence of a large observed ICC will produce a large design effect and a wide confidence interval for the treatment effect.

9 Choosing the method of analysis: cluster specific versus marginal models

Appropriate methods of analysis for individually randomised trials are based on standard methods of statistical analysis, which are well-documented[111] and do not need any further mention here. However, there are more complex issues to consider when analysing cluster randomised trials, where we have to account for groups of individuals within clusters, and this is the focus of this section. One simple approach is outlined, and two model based approaches are discussed and compared in the example given in Section 9.2.

The simplest way of analysing cluster randomised trials is to use the method of summary measures, as popularised by Matthews *et al.*[112] This gives equal weight to each cluster, irrespective of size and is a cluster specific method. It has a great deal to recommend it since it simply uses the summary statistics (mean or proportion) for each cluster and is easy to apply without specialist software. Alternative weights can also be chosen to reflect different precision in the estimates.[113] However, we cannot adjust for individual level covariates directly using this approach.

The two main approaches to analysing cluster randomised trials that enable covariate adjustment are using the cluster-specific (CS) model or alternatively the population-averaged (PA) (or marginal) model. Consider an outcome variable Y (which may be continuous or binary), with expected value μ and a generalised link of the form:

$$g(\mu) = \eta = \beta_0 + \beta_1 X + Z, \tag{1}$$

where g is a monotone function. We further assume that Z is a random variable with $E(Z) = 0$ and $\text{var}(Z) = \sigma_Z^2$, and is independent of the covariate X. Equation (1) is a suitable model for clustered data. We assume that the clusters are sampled from a larger population and the effect of any particular cluster i is to add a random effect Z_i to all the outcomes of individuals in that cluster. The variable X is the independent variable, and is assumed to be what is known as a 'fixed' effect. Thus in a trial, we could have $X = 1$ for an intervention and $X = 0$ for control. We assume that the effect of the intervention will be the same if we repeated the study. A CS model measures the effect on Y of changing X, while Z is held constant. This is a common model for longitudinal data, where it is possible to imagine, say in a cross-over trial, a treatment value changing over time. However, in a cluster randomised trial, everyone in a cluster receives the same treatment, and although a CS model can be fitted, the result can be interpreted only theoretically. The alternative is to fit a model that looks at the average effect of X over

the range of Z. In this case we fit the model

$$\eta = \beta_0^* + \beta_1^* X. \tag{2}$$

This is the so-called PA model. Here, we estimate the effect of X on Y as averaged over all the clusters Z. The errors will be correlated and we can allow for this in the fitting process described below.

Neuhaus and Jewell[114] contrast the CS models and PA models by observing that model (2) is simply model (1) with the variable Z omitted. If we assume that the coefficients for X in the two models are related by $\beta_1^* \approx B\beta_1$, where B is the bias factor, then they show that for a linear or a log–linear model $B = 1$, and so the interpretation of CS and PA models is the same. However, for a logistic link, $\mu = \text{logit}\,(P(Y = 1|X, Z))$ and

$$B = 1 - \frac{\sigma_z^2}{E(\mu)\{1 - E(\mu)\}} = 1 - \rho,$$

where ρ is the correlation of the Y's within clusters assuming $\beta_1 = 0$. Since the value of ρ in community randomised trials is usually less than 0.01, the bias in incorrectly assuming a marginal model should not be great. In this case $0 < B < 1$, so the general effect of using a PA model is to attenuate the regression coefficient. We can also see that for a logistic link the greater the variability of the random variable Z, the greater the ICC and so the greater the attenuation.

9.1 Fitting models

For continuous data we can rewrite (1) as

$$y_{ij} = \beta_0 + \beta_1 X_i + Z_i + \varepsilon_{ij} = \beta_0 + \beta_1 X_1 + \upsilon_{ij}, \tag{3}$$

where the ε_{ij} are independent random errors with variance σ^2 and we can think of the υ_{ij} as random errors which are not independent. For binary data using a logistic link, the CS logistic model is

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta X_i + Z_i, \tag{4}$$

where $\pi_{ij} = E(Y_{ij}|X_i, Z_i)$. These models can be extended to include individual specific covariates X_{ij}. The random variable Z_i is assumed independent of X_i and is usually assumed to be normally distributed with mean 0 and variance σ_Z^2. In model (4), given the Z_i, the $Y_{ij}s$ are assumed independently distributed with binomial parameter π_{ij}. We can also fit a log–linear model, which has the advantage that we get relative risks rather than odds ratios, which for prospective studies are easier to interpret. However in general, logistic models are more likely to describe data well.[115] A common assumption is to use an *exchangeable* correlation structure. This means that the variance covariance matrix of the outcome would remain the same if we changed the order of the subjects within the clusters.

There are different methods of fitting models (1) and (2) to data. The CS model (1) can be solved by maximum likelihood. The random variable Z_i is a nuisance parameter

that can be removed from the likelihood either by integration, using a technique such as Gaussian quadrature, or approximating the integral using Laplace's method and using penalised likelihood.[116] One can also adopt an empirical Bayes approach, and by specifying non-informative priors for the random components, obtain the estimates as the modes of the posterior distribution using MCMC in WinBUGS.[117]

The PA model (2) can be fitted using the Generalised Estimating Equation (GEE) method, developed by Liang and Zeger[118] in the context of longitudinal studies. It uses an iteratively reweighted algorithm to estimate the parameters and a robust method for the standard error. The advantage of the GEE method is that even if one fits the wrong correlation structure (for example, an independent error structure) the method will still produce valid estimates. However it will be less efficient and so the effect of the intervention will have higher standard errors in this case. As a general rule of thumb the method requires a minimum of 20 clusters per treatment group, although recent modifications to the fitting algorithm have shown fewer clusters can be accommodated.[119]

We can fit model (3) but not account for the correlated nature of the error terms. Provided that η_{ij} is uncorrelated with X_i then we get a valid estimate of the treatment effect but the standard error of the treatment effect is underestimated, and should be inflated. Donner and Klar[83] give a number of methods for continuous and binary outcomes, which modify the standard error associated with either the *t*-test or the chi-squared test, respectively. It is important to note that the estimate of the treatment effect is unchanged, only the standard error is inflated. An alternative method is to use the so-called 'sandwich' or Huber-White estimator, which has a long history in econometrics for estimators with continuous data and with heterogeneous variances. This fits a model where the variance of the error term $V(\eta) = \sigma_i^2$ is not assumed constant, and is estimated through the residuals of the model. The advantage of the robust standard error is that one does not need to estimate the ICC separately before conducting the analysis. It also allows for a wider range of models (such as a negative ICC which is impossible under the CS model). The disadvantage is that it may be less efficient and so would need a greater number of clusters than does use of a model which more closely reflects the actual structure.

The robust variance estimation relies on between-cluster information to ensure the validity of the resulting inferences. It is therefore important to be wary of this approach in community intervention trials, where the amount of such information tends to be relatively small. Guo *et al.*[119] discussed not only the use of a robust standard error in the Wald test, but also an alternative, the robust score test, particularly for cluster randomised trials. They show that use of the Wald test is too liberal (more likely to give a significant value) and the score test is too conservative and suggest that a simple correction $K/(K-1)$ to the score statistic will give a test size closer to the nominal value. This is relevant to primary care studies where the number of clusters is often low. The correction widens the confidence interval and so does not apply to the Wald test.

A number of papers have explored different methods of fitting models to data and have found that it often matters little which method is used provided the methods do allow for clustering.[120] With modern software there is little difficulty in using different methods. The CS methods are probably better with small numbers of clusters, provided that model (1) holds.

Table 5 Four methods of analysing the proportion of patients with an HbA1c below 7% for the DESMOND trial – intervention is structured education

Model	Odds ratio	Standard error	z	$P > \|z\|$	(95% Confidence interval)
Cluster specific	1.085539	0.166037	0.54	0.592	(0.804362, 1.465007)
Population averaged					
Robust	1.161681	0.271156	0.64	0.521	(0.735194, 1.835573)
Independent errors	1.161681	0.162818	1.07	0.285	(0.882643, 1.528933)
Exchangeable errors	1.079769	0.160086	0.52	0.605	(0.807480, 1.443876)

9.2 Example: the DESMOND trial

The Desmond trial[77] was a cluster randomised trial of structured education of 824 patients with Type II diabetes randomised into 207 clusters which were general practices. The primary outcome measure was the level of HbA1c at one year, and for this example we have dichotomised it into above or below 7%. The results of analysing the data using four of the methods available in the package *Stata 10* are shown in Table 5. For the CS method we can specify how many points are required for integration and the method of integration (we specified 60 points and adaptive quadrature). We can see that the CS method agrees well with the PA method, which assumes exchangeable correlation, although as expected the PA estimate is shrunk slightly towards the null. The robust standard error method and the PA method, which assumes independent errors gives identical estimates of the intervention effect, since the estimates are not adjusted for cluster size. However, the independence assumption is unrealistic and this method gives a different P-value to the other three methods. The robust method, which is less efficient, gives a larger standard error.

10 Discussions and conclusions

In this article we have covered the main statistical issues relevant to conducting complex interventions within a primary care setting and have placed these issues within the wider context of the MRC guidance framework.[1] Rather than attempt a fully comprehensive coverage of statistical methods applied in primary care research, let alone an exhaustive treatment of the research issues themselves, our aim in this paper has been to provide a flavour of the issues addressed over the last 7 years of the RSS Primary Health Care Study Group, that impact the role of the statistician within the research team. Moreover, the style of this paper matches closely those of the group's meetings – in particular, that of a balance between methodological issues and more practical issues of applying such methods to real-life research problems. Indeed, assuming that one accepts the phrases 'real-life' and 'research' within the same sentence, we hope that there is an implicit message emerging from the above sections – namely, that for most statisticians working in primary care, for most of their time the methodological and contextual challenges are inextricably linked.

This contention perhaps begs the question of how much of the methods covered here and in future meetings are specific to primary care research. As indicated in the relevant

sections, while few if any are unique to primary care, much of what has been described is either especially pertinent to this context, has closely linked practical issues that are specific to primary care, or both of these. Arguably much more important, though, is the question of how to ensure that the particular aspects of the design,[1,2] conduct and reporting[79] of health care research in a primary care setting can continue to attract the attention of methodologists in general and statisticians in particular.

The challenges of at least maintaining and preferably expanding the capacity of both methodological and applied expertise relevant to this agenda do of course remain. In conclusion, then, if the present paper is of some help to those currently engaged in primary care research and assists in some way in drawing those with statistical expertise to the field, then progress in achieving the Group's objectives will certainly have been made.

References

1 Craig P, Dieppe P, Macintyre S, Mitchie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. *British Medical Journal* 2008; **337**: a1655 (see also www.mrc.ac.uk/complexinterventionsguidance).

2 Campbell NC, Murray E, Darbyshire J *et al*. Designing and evaluating complex interventions to improve health care. *British Medical Journal* 2007; **334**: 455–9.

3 Glasziou P, Meats E, Heneghan C, Shepperd S. What is missing from descriptions of treatment in trials and reviews? *British Medical Journal* 2008; **336**(7659): 1472–4.

4 Mant J, McManus RJ, Hare R. Applicability to primary care of national clinical guidelines on blood pressure lowering for people with stroke: cross sectional study. *British Medical Journal* 2006; **332**(7542): 635–7.

5 Mant J, Fitzmaurice D, Murray E, Lip GY, Hobbs FD. Long term anticoagulation or antiplatelet treatment. Inclusion criteria determine results of review. *British Medical Journal* 2001; **323**(7306): 233–4; author reply 235–6.

6 Higgins J, Green S. *Cochrane handbook for systematic reviews of interventions*. Cochrane Collaboration. John Wiley and Sons, Chichester (UK); 2008.

7 Sutton AJ, Abrams KR. Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research* 2001; **10**(4): 277–303.

8 Sutton AJ, Cooper NJ, Abrams KR, Lambert PC, Jones DR. A Bayesian approach to evaluating net clinical benefit allowed for parameter uncertainty. *Journal of Clinical Epidemiology* 2005; **58**(1): 26–40.

9 Heneghan C, Perera R, Ward AA, Fitzmaurice D, Meats E, Glasziou P. Assessing differential attrition in clinical trials: self-monitoring of oral anticoagulation and type II diabetes. *BMC Medical Research Methodology* 2007; **7**: 18.

10 Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. *Annals of Internal Medicine* 2008; **149**(12): 889–97.

11 Whiting PF, Sterne JA, Westwood ME *et al*. Graphical presentation of diagnostic information. *BMC Medical Research Methodology* 2008; **8**: 20.

12 Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2007; **8**(2): 239–51. Erratum in: Biostatistics 2008; **9**(4): 779.

13 Williamson PR, Lancaster GA, Craig JV, Smyth RL. Meta-analysis of method comparison studies. *Statistics in Medicine* 2002; **21**: 2013–25.

14 Craig JV, Lancaster GA, Taylor S, Williamson PR, Smyth RL. Infrared ear thermometry compared with rectal thermometry in children: a systematic review. *Lancet* 2002; **360**: 603–9.

15 Skinner JS, Cooper A, Feder GS. Secondary prevention for patients after a myocardial infarction: summary of NICE guidance. *British Medical Journal* 2007; **334**(7603): 1112–3.
16 Feder G, Cryer C, Donovan S, Carter Y. Guidelines for the prevention of falls in people over 65. The Guidelines' Development Group. *British Medical Journal* 2000; **321**(7267): 1007–11.
17 Hippisley-Cox J, Coupland C, Vinogradova Y *et al*. Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *British Medical Journal* 2008; **336**(7659): 1475–82.
18 Burgers JS, Bailey JV, Klazinga NS, Van der Bij AK, Grol R, Feder G. Inside Guidelines. *Diabetes Care* 2002; **25**: 1933–9.
19 Grant J, Cottrell R, Cluzeau F, Fawcett G. Evaluating "payback" on biomedical research from papers cited in clinical guidelines: applied bibliometric study. *British Medical Journal* 2000; **320**: 1107–11.
20 Silagy CA, Stead LF, Lancaster T. Use of systematic review in clinical practice guidelines: case study of smoking cessation. *British Medical Journal* 2001; **323**: 833–6.
21 Lewin S, Glenton C, Oxman AD. Use of qualitative methods alongside randomized controlled trials of complex healthcare interventions: methodological study. *British Medical Journal* 2009; **339**: b3496.
22 Mays N, Pope C. Qualitative research in health care: assessing quality in qualitative research. *British Medical Journal* 2000; **320**: 50–2.
23 Harvey E, Farrin A, Underwood M, Morton V, for the UK BEAM Trial Team. What's it got to do with us? Including support staff in general practice training sessions. *Education for Primary Care* 2004; **15**: 606–10.
24 Mason VL, Shaw A, Wiles NJ *et al*. GPs' experiences of primary care mental health research: a qualitative study of the barriers to recruitment. *Family Practice* 2007; **24**: 518–25.
25 Featherstone K, Donovan JL. Random allocation or allocation at random? Patients' perspectives of participation in a randomised controlled trial. *British Medical Journal* 1998; **317**: 1177–80.
26 Campbell R, Dieppe P, Quilty B. Discrepancies between patient based assessments of outcome. *British Medical Journal* 2003: **326**; 252–3.
27 O'Cathain A, Walters SJ, Nicholl JP, Thomas KJ, Kirkham M. Use of evidence based leaflets to promote informed choice in maternity care: randomised controlled trial in everyday practice. *British Medical Journal* 2002; **324**: 643–6.
28 Stapleton H, Kirkham M, Thomas G. Qualitative study of evidence based leaflets in maternity care. *British Medical Journal* 2002; **324**: 639.
29 Moore L, Campbell R, Whelan A, Lupton P, Frohlich J, Misselbrook L. Self-help smoking cessation in pregnancy: a cluster randomised trial. *British Medical Journal* 2002: **325**: 1383–6.
30 Hill J, Foster N, Hughes R, Hay E. Meeting the challenges of research governance. *Rheumatology* 2005; **44**: 571–2.
31 Department of Health. *Research Governance Framework for health and social care*. Department of Health, London; 2005 (www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4108962).
32 National Institute for Health Research. *Research in the NHS – human resources (HR) good practice resource pack: HR good practice*. NIHR; January 2009.
33 Appleton J, Caan W, Cowley S, Kendall S. Busting the bureaucracy: lessons from research governance in primary care. *Community Practitioner* 2007; **80**; 2.
34 Welcome Trust. *Towards consensus for best practice. Use of patient records from general practice for research*. Wellcome Trust, London; 2009.
35 Williamson PR, Hutton JL, Bliss J, Blunt J, Campbell M, Nicholson R. Statistical review by research ethics committees. *Journal of the Royal Statistical Society A* 2000, **163**: 5–13.
36 Doran T, Fullwood C, Gravelle H *et al*. Pay-for-performance programs in family practices in the United Kingdom. *New England Journal of Medicine* 2006; **355**: 375–84.
37 Lancaster GA, Dodd SR, Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *Journal of Evaluation in Clinical Practice* 2004; **10**(2): 307–12.

38 Ng CJ. The development and evaluation of a Patient Decision Aid for people with Type 2 diabetes making decisions about starting insulin treatment, PhD Thesis (in preparation), University of Sheffield; 2009.

39 O'Connor AM, Jacobsen MJ. *Workbook on developing and evaluating Patient Decision Aids*, University of Ottawa, Canada; 2003.

40 Eldridge S, Spencer A, Cryer C, Parsons S, Underwood M, Feder G. Why modelling a complex intervention is an important precursor to trial design: lessons from studying an intervention to reduce falls-related injuries in older people. *Journal of Health Services Research and Policy* 2005; **10**: 133–42.

41 Lancaster GA. Statistical issues in the assessment of health outcomes in children: a methodological review. *Journal of the Royal Statistical Society A* 2009; **172**(4): 707–27.

42 Lancaster GA, Taylor S, Chellaswamy H, Lyon D, Dowrick C. Design of a clustered observational study to predict emergency admissions in the elderly: statistical reasoning in clinical practice. *Journal of Evaluation in Clinical Practice* 2007; **13**: 169–78.

43 UK BEAM Trial Team. UK Back pain Exercise And Manipulation (UK BEAM) trial – national randomised trial of physical treatments for back pain in primary care: objectives, design and intervention. *BMC Health Services Research* 2003; **3**: 16.

44 UK BEAM Trial Team. UK Back pain Exercise And Manipulation (UK BEAM) randomised trial: effectiveness of physical treatments for back pain in primary care. *British Medical Journal* 2004; **329**: 1377–8.

45 UK BEAM Trial Team. UK Back pain Exercise And Manipulation (UKBEAM) randomised trial: cost-effectiveness of physical treatments for back pain in primary care. *British Medical Journal* 2004; **329**: 1381–5.

46 Farrin A, Russell I, Torgerson D, Underwood M on behalf of UK BEAM Trial Team. Differential recruitment in a cluster randomised trial in primary care – the experience of the UK Back pain, Exercise, Active management and Manipulation (UK BEAM) feasibility study. *Clinical Trials* 2005; **2**: 119–24.

47 Garratt A. Patient reported outcome measures in trials. *British Medical Journal* 2009; **338**: a2597.

48 Lall R, Walters SJ, Campbell MJ, Morgan K, MRC CFAS. A review of ordinal regression models applied on Health Related Quality of Life Assessments. *Statistical Methods in Medical Research* 2002; **11**: 49–6.

49 Hurn J, Kneebone I, Cropley M. Goal setting as an outcome measure: a systematic review. *Clinical Rehabilitation* 2006; **20**: 756–2.

50 Hiscock H, Bayer JK, Price A, Ukoumunne OC, Rogers S, Wake M. Universal parenting programme to prevent early childhood behavioural problems: cluster randomised trial. *British Medical Journal* 2008; **336**: 318–21.

51 Foster NE, Thomas E, Barlas P *et al*. Acupuncture as an adjunct to exercise based physiotherapy for osteoarthritis of the knee: randomised controlled trial. *British Medical Journal* 2007; **335**(7617): 436–48.

52 Merbitz C, Morris J, Grip JC. Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation* 1989; **70**: 308–12.

53 Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation* 1989; **70**: 857–60.

54 Tennant A, Penta M, Tesio L *et al*. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. *Medical Care* 2004; **42**(Suppl 1): 137–48.

55 Durez P, Fraselle V, Houssiau F, Thonnard JL, Nielens H, Penta M. Validation of the ABILHAND questionnaire as a measure of manual ability in patients with rheumatoid arthritis. *Annals of the Rheumatic Diseases* 2007; **66**(8): 1098–105.

56 Rasch G. *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research, Copenhagen; 1960; Expanded edition with foreword and afterword by B.D. Wright. University of Chicago Press, Chicago; 1980.

57 Kahler E, Rogausch A, Brunner E, Himmel W. A parametric analysis of ordinal quality of life data can lead to erroneous results. *Journal of Clinical Epidemiology* 2008; **61**: 475–80.

58 Walters SJ, Campbell MJ, Lall R. Design and analysis of trials with quality of life as an outcome: a practical guide. *Journal of Biopharmaceutical Statistics* 2001; **11**(3): 155–76.
59 Snaith, RP. The hospital anxiety and depression scale. *Health Quality and Life Outcome* 2003; 1(29) (available at http://www.hqlo.com/content/1/1/29).
60 Stewart AL, Hays RD, Ware JE. The MOS Short-Form General Health Survey. Reliability and validity in a patient population. *Medical Care* 1988; **26**(7): 724–35.
61 Eldridge SE, Kerry S, Torgerson DJ. Bias in identifying and recruiting participants in cluster randomised trials: what can be done? *British Medical Journal* 2009; **339**: b4006.
62 Ross S, Grant I, Counsell C, Gillespie W, Russell I, Prescott R. Barriers to Participation in randomised controlled trials: a systematic review. *Journal of Clinical Epidemiology* 1999; **52**(12): 1143–56.
63 Mapstone J, Elbourne D, Roberts I. Strategies to improve recruitment to research studies. *Cochrane Database of Systematic Reviews* 2007; **2**: MR000013.
64 Watson J, Torgerson D. Increasing recruitment to randomised trials: a review of randomised controlled trials. *BMC Medical Research Methodology* 2006, **6**: 34, doi:10.1186/1471-2288-6-34.
65 Sullivan FM, Swan IRC, Donnan PT, Morrison JM, Smith BH, McKinstry B *et al*. Early treatment with prednisolone or acyclovir in Bell's Palsy. *New England Journal of Medicine* 2007; **357**: 1598–607.
66 McKinstry B, Hammersley V, Daly F, Sullivan F. Recruitment and retention in a multicentre randomised controlled trial in Bell's palsy: a case study. *BMC Medical Research Methodology* 2007; **7**: 15.
67 Bower P, Wilson S, Mathers N. How often do UK primary care trials face recruitment delays? *Family Practice* 2007; **24**: 601–3.
68 European Medicines Agency (EMEA). *ICH Topic E 6 (R1) Guideline for Good Clinical Practice, note for guidance on good clinical practice (CPMP/ICH/135/95)*. EMEA, London; 2006.
69 Sim I, Peterson K, Speedie S *et al*. The electronic primary care research network: a grid-based computing infrastructure for community-based clinical trials in primary care, *Journal of General Internal Medicine* 2007; **22**(Suppl 1): 59.
70 Mallen CD, Peat G, Wathall S *et al*. Recruiting patients to observational studies during the general practice consultation: findings from the PROG-RES study. *Family Medicine* 2009; **41**(Suppl 1): 22.
71 Henderson M, Wight D, Raab GM *et al*. Impact of a theoretically based sex education programme (SHARE) delivered by teachers on NHS registered conceptions and terminations: final results of cluster randomised trial. *British Medical Journal* 2007; **334**(7585): 103–4.
72 Kinmonth AL, Wareham NJ, Hardeman W *et al*. Efficacy of a theory-based behavioural intervention to increase physical activity in an at-risk group in primary care (ProActive UK): a randomised trial. *Lancet* 2008; **371**(9606): 41–8.
73 Mann E, Prevost AT, Griffin S *et al*. Impact of an informed choice invitation on uptake of screening for diabetes in primary care (DICISION): trial protocol. *BMC Public Health* 2009; **9**: 63.
74 Bankhead C, Richards SH, Peters TJ *et al*. Improving attendance for breast screening among recent non-attenders: a randomised controlled trial of two interventions in primary care. *Journal of Medical Screening* 2001; **8**(2): 99–105.
75 Little P, Turner S, Rumsby K *et al*. Dipsticks and diagnostic algorithms in urinary tract infection: development and validation, randomised trial, economic analysis, observational cohort and qualitative study. *Health Technology Assessment* 2009; **13**(19): iii–iv, ix–xi, 1–73.
76 Scott NW, McPherson GC, Ramsay CR, Campbell MK. The method of minimization for allocation to clinical trials: a review. *Controlled Clinical Trials* 2002; **23**: 662–74.
77 Davies MJ, Heller S, Skinner TC *et al*. Diabetes Education and Self Management for Ongoing and Newly Diagnosed Collaborative. Effectiveness of the diabetes education and self management for ongoing and newly diagnosed (DESMOND) programme for people with newly diagnosed type 2 diabetes: cluster randomised controlled trial. *British Medical Journal* 2008; **336**(7642): 491–5.

78 Delaney BC, Qume M, Moayyedi P *et al*. Helicobacter pylori test and treat versus proton pump inhibitor in initial management of dyspepsia in primary care: multicentre randomised controlled trial (MRC-CUBE trial). *British Medical Journal* 2008; **336**(7645): 651–4.
79 Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomized trials. *British Medical Journal* 2004; **328**: 707–8.
80 Bland JM, Kerry SM. Statistics notes. Trials randomised in clusters. *British Medical Journal* 1997; **315**(7108): 600.
81 Hewitt CE, Torgerson DJ, Miles JN. Individual allocation had an advantage over cluster randomization in statistical efficiency in some circumstances. *Journal of Clinical Epidemiology* 2008; **61**(10): 1004–8.
82 Thompson SG. The merits of matching in community intervention trials: a cautionary tale. *Statistics in Medicine* 1998; **17**(18): 2149–52.
83 Donner A, Klar N. *Design and analysis of cluster randomization trials in health research*. Arnold, London; 2000.
84 Raab GM, Butcher I. Balance in cluster randomized trials. *Statistics in Medicine* 2001; **20**(3): 351–65.
85 Feder G, Griffiths C, Highton C, Eldridge S, Spence M, Southgate L. Do clinical guidelines introduced with practice based education improve care of asthmatic and diabetic patients? A randomised controlled trial in general practices in east London. *British Medical Journal* 1995; **311**(7018): 1473–8.
86 Torgerson D, Sibbald B. Understanding controlled trials. What is a patient preference trial? *British Medical Journal* 1998; **316**: 360.
87 Majeed A, Moser K, Maxwell R. Age, sex and practice variations in the use of statins in general practice in England and Wales. *Journal of Public Health Medicine* 2000; **22**(3): 275–9.
88 Reid FD, Cook DG, Majeed A. Explaining variation in hospital admission rates between general practices: cross sectional study. *British Medical Journal* 1999; **319**(7202): 98–103.
89 Majeed A, Moser K. Age- and sex-specific antibiotic prescribing patterns in general practice in England and Wales in 1996. *British Journal of General Practice* 1999; **49**(446): 735–6.
90 Campbell MJ. Cluster randomized trials in general (family) practice research. *Statistical Methods in Medical Research* 2000; **9**(2): 81–94.
91 Doran T, Fullwood C, Kontopantelis E, Reeves D. Effect of financial incentives on inequalities in the delivery of primary clinical care in England: analysis of clinical activity indicators for the quality and outcomes framework. *Lancet* 2008; **372**(9640): 728–36.
92 Doran T, Fullwood C, Reeves D, Gravelle H, Roland M. Exclusion of patients from pay-for-performance targets by English physicians. *New England Journal of Medicine* 2008; **359**(3): 274–84.
93 Reeves D, Campbell SM, Adams J, Shekelle PG, Kontopantelis E, Roland MO. Combining multiple indicators of clinical quality: an evaluation of different analytic approaches. *Medical Care* 2007; **45**(6): 489–96.
94 Omar RZ, O'Sullivan C, Petersen I, Islam A, Majeed A. A model based on age, sex, and morbidity to explain variation in UK general practice prescribing: cohort study. *British Medical Journal* 2008; **337**: a238.
95 Sullivan CO, Omar RZ, Ambler G, Majeed A. Case-mix and variation in specialist referrals in general practice. *British Journal of General Practice* 2005; **55**(516): 529–33.
96 Littenberg B, MacLean CD. Intra-cluster correlation coefficients in adults with diabetes in primary care practices: the Vermont Diabetes Information System field survey. *BMC Medical Research Methodology* 2006; **6**: 20.
97 Elley CR, Kerse N, Chondros P, Robinson E. Intraclass correlation coefficients from three cluster randomised controlled trials in primary and residential health care. *Australian and New Zealand Journal of Public Health* 2005; **29**(5): 461–7.
98 Parker DR, Evangelou E, Eaton CB. Intraclass correlation coefficients for cluster randomized trials in primary care: the cholesterol education and research trial (CEART). *Contemporary Clinical Trials* 2005; **26**(2): 260–7.
99 Knox SA, Chondros P. Observed intra-cluster correlation coefficients in a cluster survey sample of patient encounters in general practice in Australia. *BMC*

Medical Research Methodology 2004; **4**(1): 30.

100 Adams G, Gulliford MC, Ukoumunne OC, Eldridge S, Chinn S, Campbell MJ. Patterns of intra-cluster correlation from primary care research to inform study design and analysis. *Journal of Clinical Epidemiology* 2004; **57**(8): 785–94.

101 Keogh-Brown MR, Bachmann MO *et al*. Contamination in trials of educational interventions. *Health Technology Assessment* 2007; **11**(43): iii, ix–iii,107.

102 Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *International Journal of Epidemiology* 2006; **35**(5): 1292–300.

103 Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *International Journal of Epidemiology* 2006; **35**(5): 1292–300.

104 Kerry SM, Bland JM. Unequal cluster sizes for trials in English and Welsh general practice: implications for sample size calculations. *Statistics in Medicine* 2001; **20**: 377–90.

105 Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. *Clinical Trials* 2004; **1**(1): 80–90.

106 Turner RM, Thompson SG, Spiegelhalter DJ. Prior distributions for the intracluster correlation coefficient, based on multiple previous estimates, and their application in cluster randomized trials. *Clinical Trials* 2005; **2**(2): 108–18.

107 Turner RM, Omar RZ, Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. *Statistics in Medicine* 2001; **20**(3): 453–72.

108 Eldridge S, Ashby D, Bennett C, Wakelin M, Feder G. Internal and external validity of cluster randomised trials: systematic review of recent trials. *British Medical Journal* 2008; **336**(7649): 876–80.

109 Puffer S, Torgerson D, Watson J. Evidence for risk of bias in cluster randomised trials: review of recent trials published in three general medical journals. *British Medical Journal* 2003; **327**: 785–9.

110 Turner RM, Prevost AT, Thompson SG Allowing for imprecision of the intracluster correlation coefficient in the design of cluster randomised trials. *Statistics in Medicine* 2004; **23**(8): 1195–214.

111 Chow S-C, Liu J-P. *Design and analysis of clinical trials: concepts and methodologies*, 2nd edn. John Wiley and Sons Inc, Malden, USA; 2004.

112 Matthews JNS, Altman DG, Campbell MJ and Royston P. Analysis of serial data using summary measures. *British Medical Journal* 1990; **300**: 230–5.

113 Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M. Analysis of cluster randomized trials in primary care: a practical approach. *Family Practice* 2000; 17(2): 192–6.

114 Neuhaus JM, Jewell NP. A geometric approach to assess bias due to omitted covariates in generalized liner models. *Biometrika* 1993; **80**: 807–15.

115 Campbell MJ. Should we use relative risks or odds ratios in cluster randomised trials with binary outcomes that have high proportions? *Journal of Epidemiology and Community Health* 2008; **62**(Suppl 1): A24.

116 Campbell MJ, Donner A, Klar N. Developments in cluster randomized trials and statistics in medicine. *Statistics in Medicine* 2007; **26**: 2–19.

117 Spiegelhalter DJ. Bayesian methods for cluster randomized trials with continuous response. *Statistics in Medicine* 2001; **20**: 435–52.

118 Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**: 13–22.

119 Guo X, Pan W, Connett JE, Hannan PJ, French SA. Small-sample performance of the robust score test and its modifications in generalized estimating equations. *Statistics in Medicine* 2005; **24**: 3479–95.

120 Peters TJ, Richards SH, Bankhead CR, Ades AE, Sterne JAC. Comparison of methods for analysing cluster randomized trials: an example involving a factorial design. *International Journal of Epidemiology* 2003; **32**: 840–6.