1

Introduction

In this chapter, we will discuss the rationale for randomised trials and how cluster trials differ from individually randomised trials. The development of cluster trials and how they fit into the framework of complex interventions will be outlined. We will describe a number of trials that will be discussed throughout the book. Two fundamental concepts, namely, the unit of inference and how to measure the degree of clustering will also be discussed.

1.1 Randomised controlled trials

How do we know that a treatment works? It has long been asserted that the only way of assessing whether a treatment actually works is through a randomised controlled trial (RCT). *Testing Treatments*, an excellent book (Evans *et al*., 2011, available free at www.testingtreatments.org), gives a series of examples where treatments, thought to be beneficial on the basis of observational data, have been shown, in fact, to harm patients. The modern paradigm is the example of hormone replacement therapy, which had been perceived as beneficial until the Women's Health Initiative trial (Prentice *et al*., 1998) and other studies showed that, far from reducing the risk of heart disease, it actually slightly increased the risk.

The main ingredients of an RCT can be labelled as 'ABC' (Campbell, 1999).

1.1.1 A-Allocation at random

This means that who gets the new treatment, which is to be evaluated, and who does not is determined by chance. These days this usually means allocation is determined by a computer-generated random sequence. However, in the past, this was done with shuffled

How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research,
First Edition. Michael J. Campbell and Stephen J. Walters.

Companion website: http://www.wiley.com/go/cluster_randomised

envelopes and other mechanical means such as tossing a coin. The main purpose of randomisation is to ensure that, in the long run, the only consistent difference between the randomised groups is that one group got the new treatment and the other did not; all other differences have been averaged out. Factors that might influence outcome are often called *prognostic factors*. For example, people with more severe disease at the start of treatment may be expected to do worse than people with mild disease. The important point about randomisation is that it ensures, in the long run, that there is no preponderance of a prognostic factor in one group compared with another. A further point is that this is true for both known and unknown factors. Thus if, after a trial had been published, it became known that a certain gene had prognostic significance, even though it would be too late to measure the gene in the patients, the investigators are protected from major imbalances in the gene frequency in the treatment and control groups by randomisation. An operative phrase here is 'in the long run'. Trials cannot be infinitely large, and so for any trial of finite size, it may be possible to find imbalances in prognostic factors, and steps may be needed to control these. As we shall see later, cluster trials are primarily judged on the number of clusters they contain and since this is often not large, imbalance is a particular problem.

Simple randomisation means that the treatment allocation is determined purely by chance. However, this might mean that the numbers in each group are unequal, which usually reduces the efficiency of the study. Thus, a development is *blocked* randomisation, whereby an even number of subjects are selected and randomised so that half of the subjects get one treatment and the other half the alternative treatment.

If there are known important prognostic factors in a trial, then it would be foolish to leave a balanced outcome to chance, and so *stratified* randomisation is carried out. Here, the subjects are divided into groups or strata depending on the prognostic factor and blocked randomisation carried out within each stratum. For example, patients might be divided into those with severe disease and those with mild disease, and then randomisation carried out separately within those two disease severity groups. This ensures that there are approximately the same number of patients in each treatment group with severe disease and the same number with mild disease.

Another important feature of randomisation is that neither the patient nor the person recruiting the patient knows in advance which treatment the patient is to receive.

1.1.2 B-Blindness

This means that the treatment is concealed to either the investigator or the patient. A *double-blind* trial means that neither the investigator nor the patient knows which treatment they are getting. Blindness can be important because belief can prove an important part in a patient's recovery and outcome. In some cases, such as whether to plaster a fracture or not, it may be impossible to blind the patient. However, it may still be possible to blind the person measuring the outcome of the trial.

1.1.3 C-Control

This usually refers to contemporaneous controls. This means that patients are evaluated at the same time in each group. Other factors which affect all patients, such as improved quality of care, should affect the intervention group and the control group equally. The control may comprise 'treatment as usual' (sometimes abbreviated to tau), which is common for non-pharmacological treatments, or a placebo for pharmacological treatments. A *placebo* is an inert compound that physically resembles the active drug, so that the patient is unaware

whether they have taken the drug with the active compound. They are used because often the very act of giving treatments will bring about improvements, irrespective of the actual treatment. An alternative control is another active treatment. Usually it is helpful to know in advance that this active treatment is effective relative to no treatment, because an inconclusive result (i.e. no difference between the two treatments on test) would mean that we would be unable to decide if the new treatment was beneficial or not relative to no treatment.

The idea of testing treatments has been shown by many authors to have a long history. However, it was not until the 1940s that trials that used proper randomisation, contemporaneous controls and blindness were published (MRC, 1948). These initial trials were individually randomised and analysed, that is individual patients were randomised to alternative treatments and then the outcome was measured on these patients. This has formed the gold standard for assessing medical treatments ever since.

1.2 Complex interventions

However, often interventions are not single simple interventions such as drugs, but so-called complex interventions, with a variety of interacting components. The United Kingdom's Medical Research Council website has a good description of these (http://www.mrc.ac.uk/Utilities/Documentrecord/index.htm?d=MRC004871).

Examples of complex interventions might include specialist stroke units, training surgeons in a new technique and a leaflet campaign to get children to take their asthma medication. Here, a number of patients will all be treated in the same unit; for example a surgeon will operate on a number of different patients and so all these patients will benefit (or suffer!) from the same level of skill, namely that possessed by that surgeon.

Let us consider an example of an RCT to evaluate the clinical effectiveness of a new surgical technique compared to existing surgical techniques in a population of patients undergoing surgery. There are at least three possible designs for a proposed RCT to evaluate a new surgical technique:

Design 1: All surgeons are trained in the new technique. When a patient presents for surgery to a particular surgeon, the surgeon is told, via a randomisation method, which type of surgery to use.

Design 2: Some surgeons are already trained in the new technique and some are not (perhaps they had to volunteer for training). If a patient presents who is eligible for surgery, they are randomised to either a surgeon using the new technique or the one using the old.

Design 3: Willing surgeons are randomised to be either trained in the new technique or not. Those trained in the new technique will then use it when appropriate. Patients arrive at surgery through the usual channels.

Each of these designs has advantages and disadvantages. In Design 1, the randomisation is conducted *within* a surgeon, so we can compare patients operated on by the same surgeon with the new technique against those treated with the standard method. Thus, the fact that some surgeons are better and more experienced than others will not affect the comparison. On the other hand, it may be very difficult for a surgeon who has been trained in a new technique to revert to a former mode of practice, and we do not know if the training might have improved the outcomes for the standard method as well.

In Design 2, patients treated by a good and experienced surgeon could be expected to do better than patients treated by an inexperienced (poor) surgeon, so the comparisons will depend not just on the patient but also on the surgeon. These are so-called *therapist* trials and are a form of cluster trial which we will discuss later. It is also possible that the better surgeons are the ones who volunteer for further training, so confounding the effects of experience and the new technique.

In Design 3, we have a truly randomised comparison. However, again each surgeon can be expected to treat a number of patients, and these patients' outcomes will be affected by the surgeons' skill, training and experience. Thus, the outcomes from the patients are not completely independent. These are *cluster randomised controlled* trials (cRCTs).

Thus, a cluster randomised trial is a trial in which groups of subjects are randomised rather than individuals. They are sometimes known as *group randomised trials*. The key important fact is that outcomes for subjects in one cluster are not independent. This means that conventional methods of statistical analysis which typically assume independence, of outcomes, are invalid and likely to give incorrect results.

1.3 History of cluster randomised trials

The first paper to recognise explicitly the issues that arise when groups of subjects are randomised was by Cornfield (1978). He originated the sayings:

> Randomisation by cluster accompanied by an analysis appropriate to randomisation by individual is an exercise in self-deception

and

> Analyse as you randomise.

The latter says that if randomisation was by cluster, then analysis should be by cluster.

Cornfield's paper was followed by the pioneering work of Allan Donner who has written numerous papers on the subject since 1981. There have also been a number of books on cluster randomised trials by Murray (1998), Donner and Klar (2000), Hayes and Moulton (2009) and Eldridge and Kerry (2012). There have also been a number of reviews of statistical methodology, for example Campbell, Donner and Klar (2007).

In the past, cluster trials were often misunderstood and poorly analysed. For example, Simpson, Klar and Donner (1995) reviewed primary prevention trials published between 1990 and 1993 and showed that out of 21 articles only 19% (4/21) included sample size calculations or discussions of power that allowed for clustering, while 57% (12/21) took clustering into account in the analysis. An important landmark was the publication of the Consolidation of Standards for Reporting Trials (CONSORT) statement for cluster trials (Campbell, 2004; Campbell, Elbourne and Altman, 2004). Sadly the evidence is that the reporting of cluster trials has not markedly improved since that statement (Ivers *et al*., 2011).

1.4 Cohort and field trials

There are two parameters that control the total size of a cluster trial: these are the number of clusters (k) and the number of subjects per cluster, the cluster size (m). The notation appears to

Table 1.1 Basic cluster trial types.

		Number of clusters (k)	
		Large	Small
Number of subjects per cluster (m)	Large	Rarely seen – very expensive!	*Field trials*, or community interventions
	Small	*Cohort trials*	*Pilot studies*

originate from Allan Donner and is quite universal now (Donner and Klar, 2000). As shown in Table 1.1, for a fixed sample size, you can have small number of clusters k with a large number of patients per cluster m, or vice versa. It is unusual to have either large k and m, or small k and m.

A trial with a small number of clusters each with a large number of subjects is often termed a *field trial*. It is common when one is trying to evaluate community-wide interventions. Such a trial is sometimes called a *cluster–cluster* trial since inference is on the way the intervention has changed a cluster level response. The trial with a small cluster size and large number of clusters is often called a *cohort trial*, since patients are followed up as a cohort or as a group.

We have found that the best way to teach and explain a concept is to start by giving some examples, so we now give some examples of field and cohort trials, so that the reader can get the idea of the range of their application.

1.5 The field/community trial

The main emphasis in field or community trials is at the cluster level. The investigator is interested in how a whole community changes its behaviour. As discussed later, this tends to lead to *cluster level* interventions. Field trials usually employ a cross-sectional sample of communities before and at different times after an intervention. In the control group, the timings of the sampling are usually the same as for the intervention group. In general, because communities are usually sampled, they do not have the problem of dropouts which are a problem with conventional individually randomised trials. It could be argued that they are more generalisable than conventional trials since they represent a random sample from the population. However, some of the sample may have recently arrived in the population and so not received the intervention. This will reduce the size of the contrast between the intervention clusters and the control clusters.

1.5.1 The REACT trial

An example of a field trial is the REACT trial (Hedges *et al*., 2000). The objective here was to determine the impact of a community educational intervention to reduce patient delay time on the use of reperfusion therapy for acute myocardial infarction (AMI). The intervention was designed to enhance patient recognition of AMI symptoms and encourage early emergency department (ED) presentation with resultant increased reperfusion therapy rates for AMI. The study took place in 44 hospitals in 20 pair-matched communities in five US geographic regions. Eligible study subjects were non-institutionalised patients without chest injury (aged >30 years) who were admitted to participating hospitals and who received a

hospital discharge diagnosis of AMI; $n = 4885$. The applied intervention was an educational programme targeting community organisations and the general public, high-risk patients and health professionals in target communities. The primary outcome was a change in the proportion of AMI patients receiving early reperfusion therapy (i.e. within 1 hour of ED arrival or within 6 hours of symptom onset). Four-month baseline was compared with the 18-month intervention period. Of the patients included in the primary analyses, 28.3% received reperfusion therapy within 6 hours of symptom onset in the intervention community group during the baseline period, compared with 27.9% in the control community group. The authors concluded that community-wide educational efforts to enhance patient response to AMI symptoms may not translate into sustained changes in reperfusion practices.

1.5.2 The Informed Choice leaflets trial

The field trial that we will describe in detail throughout the book is the Informed Choice leaflets trial (O'Cathain *et al*., 2002). The basics of this trial are given in Table 1.2. An important feature of this trial is that women were sampled before the intervention was delivered to the maternity units and again afterwards. The women were not the same in the before and after groups.

In this trial, the entire cluster is given the intervention, that is all women got the intervention and all women were given a questionnaire to measure the outcome. It should be noted

Table 1.2 Example of a field trial (O'Cathain *et al*., 2002).

Population	Samples of antenatal (at 28-wk gestation) and postnatal (8 wk after delivery) women
Intervention	Provision of 10 pairs of *Informed Choice* leaflets for service users and midwives and a training session for staff in their use
Comparator	Usual care
Unit of randomisation	Ten maternity units (five intervention, five control) between 600 and 900 women completed questions either before or after birth, also before or after the intervention (different women in each case)
Method of randomisation	Clusters were paired on the basis of their annual numbers of deliveries to ensure balance in the numbers in the two arms of the trial. Pairs were randomly assigned by tossing a coin to receive the set of leaflets (five intervention units) or to continue with usual care (five control units)
Outcome	Binary primary outcome – proportion of women who answered 'yes' to the question '*Have you had enough information and discussion with midwives or doctors to make a choice together about all the things that happened during maternity care*?'
Data collection	Through postal questionnaires

that the method of randomisation described, namely the toss of a coin, is not recommended because it cannot be replicated or verified.

1.5.3 The Mwanza trial

Grosskurth *et al*. (1995) cited in Hayes and Moulton (2009) described a trial whose aim was to reduce the prevalence of human immunodeficiency virus (HIV) infection by treating other sexually transmitted diseases (STDs) in the rural Mwanza region of Tanzania. HIV incidence was compared in six intervention communities and six pair-matched comparison communities. This is described in Table 1.3.

A total of 12 537 individuals were recruited. At the follow-up, 8845 (71%) of the cohort were seen. There was concern that a control community might be affected by having an intervention community adjacent to it, so an attempt was made to separate the control and intervention communities. Note that in contrast to the Informed Choice leaflets trial, in this case the investigators did not want to include all the subjects and so they chose a random sample of subjects. Also in contrast to the Leaflets trial, the same subjects were followed up before and after the intervention.

1.5.4 The paramedics practitioner trial

An example where the cluster is a period of time is given by a study described by Mason *et al*. (2007) which is a cluster randomised trial of paramedic practitioners. The intervention was delivered by seven paramedics who were trained to provide community assessment and treatment of patients aged over 60 who contacted the emergency ambulance services between 8 a.m. and 8 p.m. The outcome variables were ED attendance or hospital admission within 28 days of call, and satisfaction with service. The randomisation was to different time periods which were weeks when paramedic practitioners are either operative or non-operative. In total,

Table 1.3 The Mwzana trial (Grosskurth *et al*., 1995).

Population	Adults aged 15–54 yr living in the Mwanza region of Tanzania
Intervention	The establishment of an STD reference clinic, staff training, regular supply of drugs, regular supervisory visits to health facilities and health education about sexually transmitted disease
Comparator	Usual care
Unit of randomisation	Communities
Method of randomisation	Twelve communities were paired by geography, type of settlement and prior STD attendance. The method of randomisation is not stated
Outcome	HIV incidence
Data collection	A random cohort of about 1000 adults aged 15–54 from each community was surveyed at baseline and at follow-up 2 yr later

Table 1.4 The paramedics practitioner trial (Mason *et al*., 2007).

Population	Patients aged over 60 who called the emergency services with minor acute conditions
Intervention	Seven experienced paramedics were given a training course to enable them to provide community-based clinical assessment and were included in the operation of the ambulance service
Comparator	Weeks when the paramedics were removed from operational duties
Unit of randomisation	Weeks when the paramedics were on duty or not
Method of randomisation	Weeks were randomised before the start of the study to the paramedic practitioner service being active (intervention) or inactive (control), when the standard 999 service was available. Randomisation by blocks, with a block size of 6, with 26 weeks in the intervention and 30 weeks in the control
Outcome	Attendance at emergency department and hospital admission between 0 and 28 days, interval from time of call to time of discharge and patients' satisfaction with the service received
Data collection	The emergency department or ambulance service records were used to collect clinical data, including investigations, treatment, diagnoses and discharge from the service, relating to the initial patient episode

1549 individuals were recruited to the intervention over 26 weeks and 1469 individuals were recruited to the control over 30 separate weeks. The details are given in Table 1.4.

1.6 The cohort trial

The second type of cluster trial is closer in design to an individually randomised trial and typically uses more clusters and relatively smaller cluster sizes than a field trial. Usually the same patients are followed up over time and it has been termed a *cohort trial*. The main emphasis in a cohort trial is the individual. Usually the investigator is interested in looking at change in an outcome, and so will measure values at baseline and at various times after the intervention has taken place. A cohort cluster trial suffers from similar problems of generalisation as a conventional trial. Patients have to be willing to be in the trial and may drop out before follow-up. If only a small number of people approached to enter the trial volunteer, or if a large number of people drop out before follow-up, then the generalisability of the trial could be called into question. Thus, it is important to report dropout rates and do sensitivity analyses to consider whether the nature of the dropouts may affect the conclusions.

1.6.1 The PoNDER trial

An example of a cohort cRCT that will be used throughout the book is the Postnatal Depression Economic Evaluation and Randomised Controlled Trial (PoNDER) study.

Table 1.5 The PoNDER study (Morrell *et al*., 2009).

Population	Post-natal women
Intervention	Health visitors trained to providing psychologically informed sessions based on CBA approaches or PCA approaches for an hour a week for 8 wk to new mothers with PND
Comparator	Usual care
Unit of randomisation	101 general practices and their associated health visitors 63 in intervention arms (30 CBA and 32 PCA) and 39 in the control
Method of randomisation	Clusters were stratified by the number of expected births per cluster per year into three groups (<70, 70–100 and >100). Clusters were allocated to either CBA or PCA, the two interventions or the control group in a ratio of 1 : 1 : 1
Outcome	The main outcome measure was the 6-month EPDS score. The EPDS is scored on a 0–30 scale with a higher score indicating more depressive symptoms. A score of 12 or more is regarded as being 'at higher risk' of PND
Data collection	Women were sent a postal questionnaire at 6 weeks post-natally to collect demographic details, measure depressive symptoms using the EPDS and measure social support and stressful life events using the measure of social relationships and list of threatening experiences, and previous depression

CBA, cognitive behavioural approach; PCA, person-centred approach; (PND), postnatal depression; EPDS, Edinburgh Postnatal Depression Scale.

(Morrell *et al*., 2009). The PoNDER study involved health visitors (HVs) who worked in general practitioner (GP) practices which were randomised for the HVs to be trained or not in psychological approaches to identify postnatal depressive symptoms and to treat women with postnatal depression. The HVs sequentially recruited new mothers who collectively formed the corresponding cluster. The PoNDER trial randomised 101 clusters (GP practices and their associated HVs) and collected data on 2659 new mothers with an 18-month follow-up (Table 1.5). This trial is further described in Chapter 8, where we will introduce the patient flow diagram, which shows the number of patients available at each stage of the trial, and in particular how many women responded.

1.6.2 The DESMOND trial

Another trial which we will return to is the DESMOND (Diabetes Education and Self-Management Ongoing and Newly Diagnosed) trial (Davies *et al*., 2008), which is described in Table 1.6.

In the United Kingdom, diabetes is usually treated in primary care, and it was deemed impossible to randomise people in the same practice to different treatments. Thus, practices were chosen (at random) as either 'intervention' practices or 'control' practices. Practices randomised to deliver the DESMOND educational intervention taught the course to groups of eight people at the same time.

Table 1.6 The DESMOND trial (Davies *et al*., 2008).

Population	Patients with Type II diabetes
Intervention	A structured group education programme for 6 hours delivered in the community by two trained healthcare professional educators the 'DESMOND' course to improve lifestyle
Comparator	Usual care
Unit of randomisation	General practices (105 in intervention and 102 in the control)
Method of randomisation	Randomisation was at practice level, with stratification by training status and type of contract with the primary care organisation (General Medical Services or Personal Medical Services)
Outcome	Glycosylated haemoglobin (HbA1c%) at 1 yr
Data collection	Outcome measures were collected at baseline and at 4, 8 and 12 months. Biomedical data were collected at practice visits. Questionnaire data were collected from participants at the beginning of the study and by postal questionnaire at 4, 8 and 12 months.

1.6.3 The Diabetes Care from Diagnosis trial

A further study for which data are available is the Diabetes Care from Diagnosis trial (Kinmonth *et al*., 1998), which is described in Table 1.7. The investigators randomised GPs into those who would receive training in 'patient-centred care' and those who did not. A total of 21 practitioners were trained and 20 acted as controls. It would be difficult or impossible for a doctor to change from 'patient centred care' to 'paternalistic' care with successive patients. The outcome was measured by HbA1c% in their diabetic patients. An important distinction in this trial is that the patients were diagnosed with diabetes by the GPs during the trial. Thus, the investigators could not say in advance exactly how many patients would be in each cluster. This is in contrast to a trial of a new intervention where all the patients are present at the start of the trial.

Table 1.7 Diabetes care from diagnosis trial (Kinmonth *et al*., 1998).

Population	Patients with Type II diabetes
Intervention	Patient-centred care training (1.5 days)
Comparator	Routine care
Unit of randomisation	General practices (21 in intervention and 20 in the control)
Method of randomisation	Practices were randomised, stratified by size of practice(large or small) and method of delivery of diabetes care (specialist nurse or not)
Outcome	Body mass index at 1 yr
Data collection	Body mass index and HbA1c% were measured 1 yr after the patient had been recruited

Table 1.8 The REPOSE (Relative Effectiveness of Pumps Over multiple dose injections and Structured Education) trial.

Population	Patients with Type I diabetes
Intervention	Insulin pumps
Comparator	Multiple dose injections (MDI)
Unit of randomisation	Training courses
Method of randomisation	Courses were randomised in pairs to pumps or MDIs, within centre
Outcome	HbA1c%
Data collection	HbA1c collected at 1 yr after course

1.6.4 The REPOSE trial

The REPOSE (Relative Effectiveness of Pumps Over multiple dose injections and Structured Education) trial (REPOSE trial protocol www.sheffield.ac.uk/scharr/sections/dts/ctru/repose) is designed to examine whether insulin pumps give better control of diabetes over multiple dose injections (MDIs). Patients are trained in groups of 6–8 to use either pumps or MDIs. The outcome was HbA1c% at 1 year. To reduce recruitment bias (see later for more discussion), patients in each centre were recruited to one of the two courses and when the courses were full, randomisation was done so that one course got the intervention and the other got the control. A total of about 280 patients were to be recruited with about 20 pairs of pump/MDI courses (Table 1.8).

1.6.5 Other examples of cohort cluster trials

In the trial described by Puder *et al*. (2011), the intervention was the class in school but the randomisation was by school. They wished to test a multi-dimensional culturally tailored lifestyle intervention to improve fitness in school children. A total of 40 preschool classes in 30 school areas with a high migrant population in the German- and French-speaking regions of Switzerland were randomised (20 per group) to the intervention or control. A total of 652 of the 727 preschool children gave informed consent. The main outcome measure was aerobic fitness as measured by a shuttle run test. The same children were measured before and after the intervention.

Soncini *et al*. (2007) looked at the survival of amalgam versus composite fillings in teeth and randomised 267 children aged 6 to 10 to have amalgam fillings and 267 to have composite fillings. It was deemed simpler to ensure that each child had either amalgam or composite fillings, and so survival times of the filling were clustered by mouth. In permanent teeth, the replacement rate was 14.9% of composites versus 10.8% of amalgams, and the repair rate was 2.8% of composites versus 0.4% of amalgams. They concluded that composite restorations on posterior tooth surfaces in children may require replacement or repair at higher rates than amalgam restorations, even within 5 years of placement.

1.7 Field versus cohort designs

A summary of the differences between field trials and cohort trials is given in Table 1.9.

Table 1.9 Contrasts between field trials and cohort trials.

	Field or community trials	Cohort trials
Emphasis	Effect of an intervention at a cluster level	Effect of an intervention at the individual patient level
Generalisability	If response rate is high, then likely to be generalisable	Requires a high follow-up rate
Analysis	Cluster level	Individual level, allowing for clustering. Often will use baseline data as covariates

Because responses within the same subject often have a strong positive correlation, one can use the baseline measurement as a covariate and usually this will reduce the standard error of the treatment effect. Thus in theory, a cohort design may be more efficient than a cross-sectional design. However, Feldman and McKinlay (1992) presented a unified statistical model that embraces both designs as special cases, thus allowing an assessment of how the values of different design parameters affect their relative precision. A principal conclusion from their investigation was that cohort designs have unique disadvantages that may outweigh any advantage in theoretical efficiency. The first of these is related to possible instability in cohorts of large size, with the resulting likelihood of subject loss to follow-up. Although this disadvantage can be compensated for by oversampling at baseline, this might well negate the original reasons for adopting a cohort design. Differential loss to follow-up by intervention group also creates the risk of bias. The second disadvantage is related to the issue of representativeness of the target population, which is invariably hampered by the ageing of the cohort over time. Assuming that changes related to the ageing process are independent of the intervention assignment, this effect will not invalidate the principal comparison of interest. However, it does imply that a difference observed in a cohort trial with respect to a given outcome variable cannot be directly compared to the corresponding difference between observed cross-sectional samples. Thus if the primary questions of interest focus on change at the community level rather than at the level of the individual, cohort samples are the less natural choice. This point was discussed by Ukoumunne and Thompson (2001) and Nixon and Thompson (2003) who described and compared several approaches that might be taken to the analysis of repeated cross-sectional samples.

1.8 Reasons for cluster trials

The five main reasons for adopting a cluster randomised trial are summarised in Table 1.10.

The first reason for using a cluster trial is fear of *contamination*. This occurs when subjects in the control group are exposed to the intervention. In field trials, this is fairly obvious. People living in the same community are unlikely not to notice a mass education programme delivered on the television or local newspaper! However, in cohort trials the effect is more subtle. Doctors trained in a new technique will find it difficult to revert to an old technique at the

Table 1.10 Reasons for adopting a cluster randomised controlled trial rather than an individually randomised controlled trial.

1. Contamination
2. Reflects real-life delivery
3. Convenience
4. Intervention better delivered to groups
5. Ethics

toss of a coin and so may not deliver the standard treatment as they used to do before being trained to deliver the new treatment. The HOOP (Hands On Or Poised) trial (McCandlish *et al*., 1998) was a trial of midwives trained to have their hands either on the perineum or poised (i.e. only touching if absolutely necessary) while delivering a baby. For each delivery, they were given a random allocation as to which method they should use. Monitors had to be employed to ensure that the midwives complied with their allocated intervention. However, it could be argued that this is not how midwives usually perform. Clinicians do not usually make random choices when faced with similar patients, and so allowing a clinician to treat similar patients in a similar way ensures that the trial more closely follows real life.

Another reason is that it may be cheaper and more convenient to deliver an intervention to a group. In the Informed Choice leaflets trial, it was much easier to deliver leaflets to all the intervention clusters and ensure that they were given out to all expectant mothers. This was also true of the DESMOND trial, where it was much cheaper to deliver the intervention in groups. It is often easier to deliver an intervention to a group of people when it may involve expensive pieces of equipment which require training in their use such as in the REPOSE trial. Also it can be difficult if people notice that others are getting a different treatment. In the DESMOND trial, patients may wonder why other people with the same doctor were getting different treatment, and demand the same for themselves. It can be more effective to deliver educational programmes to groups. Patients in the same education programme will interact with each other and may learn more than if learning on their own. This was particularly true with DESMOND, where patients learned from each other as well as from the trainer.

Finally, there are ethical considerations associated with cluster trials. These have been considered by Weijer *et al*. (http://crtethics.wikispaces.com) among others. Further discussion of ethics will be given in Chapter 7, where it will be necessary to include ethical considerations in the protocol of a study.

Having said this, cluster trials are not a universal panacea, and they are more complicated and larger than conventional individually randomised trial. Puffer, Torgerson and Watson (2003) argue that the problems of recruitment bias are so severe that in many cases individual randomised trials might be better. A useful example is given by Gilbody *et al*. (2008). They examined whether cluster randomised trials produced baseline imbalances between intervention and control conditions; gave results that are substantially different from individually randomised trials and gave different results when adjusted for unit clustering. They used 14 cluster randomised trials and 20 individualised trials of the same intervention which was

collaborative care for depression. They found no baseline imbalances in either cluster randomised or individually randomised studies. Cluster randomised studies gave almost identical estimates of effect size when compared to individually randomised studies. Their conclusion was that the additional effort and expense involved in cluster randomised trials need to be justified when individualised studies might produce robust and believable results.

1.9 Between- and within-cluster variation

It is important to understand the concepts of within- and between-cluster variation and this is perhaps best illustrated with an example. In the DESMOND trial of diabetic patients, the primary outcome was the HBA1c, in percent, level at 1-year follow-up. Figure 1.1 shows the HBA1c outcome for patients in each practice, at 1-year follow-up, by intervention and control groups from the DESMOND trial (Davies *et al*., 2008) and by practice. We are interested in the difference in the mean 1-year HbA1c between the intervention and control practices. However, one can see that there is a good deal of variation within and between the practices, with some practices having, in general, high values and some having low values. This illustrates the key point that we cannot think of the outcomes for individuals as being independent, we need to *allow* for the fact that two people in the same practice are more similar than two people selected at random from different practices. Thus, there is variability, in outcomes, *between* practices and variability *within* practices. The R code to generate this figure is given in Appendix 1.A.3.

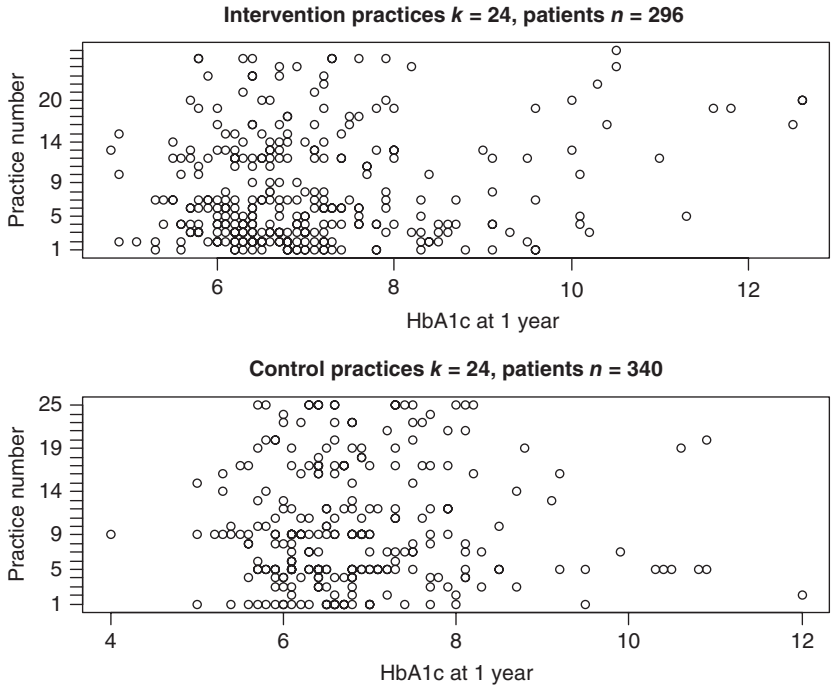

Figure 1.1 HbA1c (%) at 1 year by intervention/control from DESMOND (Davies *et al*., 2008).

1.10 Random-effects models for continuous outcomes

1.10.1 The model

Suppose we have a continuous outcome, y_{ij}, for the ith patient in the jth cluster in the cRCT. A basic linear model assuming continuous outcomes is

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_j + \epsilon_{ij} \tag{1.1}$$

where $j = 1, \ldots, k$ indexes the clusters and $i = 1, \ldots, m_j$ indexes the subjects within clusters; x_{ij} is an indicator variable for the experimental group ($0 =$ control and $1 =$ intervention); β_0 is the mean outcome in the control group and β_1 is the intervention effect; ϵ_{ij} is random error term with variance σ^2_ϵ and correlation $\text{Corr}(\epsilon_{ij}, \epsilon_{kj}) = 0$; $(i \neq k)$ α_j is a random effect of cluster j across all patients with variance σ^2_α. We assume that α_j and ϵ_{ij} are independent random variables with mean zero. This is known as a *random-effects model* or the *random intercept model*. It was first described by Eisenhart (1947).

In statistics, a random-effects model, sometimes called a *variance components model*, is a hierarchical or multilevel model of parameters that vary at more than one level. Such models are particularly appropriate for cluster RCTs and research designs where data for participants are organised at more than one level (i.e. nested data). The units of analysis are usually individuals (at a lower level) who are nested within clusters (at a higher level).

A random-effects model inference is made conditionally on the clusters using a single equation to model the covariate effects and the correlation between observations. Comparisons are made within a cluster rather than across all patients providing a treatment effect with a cluster-specific interpretation. In a random-effects model, a separate effect is fitted for each cluster in either a linear or logistic regression model. The variance of the outcome is separated into within- and between-cluster components. This provides a measure of how much of the variability in the outcome between patients can be explained by the cluster which they are in and also enables the heterogeneity within the data to be more fully explored.

The basic model (1.1) is built upon a number of assumptions. The main one is that we are assuming that each cluster is sampled from a population of clusters, and the results can be generalised to this population. It is commonly assumed that α_j and ϵ_{ij} are Normally distributed, but this is not necessary for some of the estimation procedures. The basic model assumes that the random variables have constant variance, but again this can be relaxed. Variation in α_j induces variation in the mean outcome across the clusters, represented by σ^2_α, the between-cluster variance. It assumes that the treatment effect is homogeneous across the clusters. This is a realistic model for any clustering across both treatment arms in which the clustering is unlikely to have an impact on the treatment effect.

A random-effects model can be contrasted with a fixed-effects model. In a fixed-effects model, we estimate separate parameters for each cluster. Normally a fixed effect is something important, such as a treatment effect and we assume that if we repeat the trial the same treatment effect would be seen. However, in a random-effects model, the clusters are not important, and if we repeat the trial we would not use the same clusters. Thus, a researcher investigating the effect of patient-centred care for people with diabetes administered by GPs would expect a similar effect for the intervention for two trials conducted separately in the United Kingdom. However, the trials would not use the same GPs and so simply putting dummy variables in the model for GPs would artificially reduce the variability since it would suggest we knew which

GPs were to be used in the second trial. One can see from model (1.1) that the variance of the outcome is

$$V(y_{ij}) = \sigma_\alpha^2 + \sigma_\epsilon^2$$

and so that the effect of the clusters if this model pertains is to *increase* the variability of the outcome.

1.10.2 The intracluster correlation coefficient

The parameter σ_α^2 is the *between-cluster* (*or intracluster*) variance and σ_ϵ^2 is the within-cluster variance. For a basic model, we assume that all the k clusters have the same size so that $m_j = m$ for all j. Another generalisation is to allow σ_α^2 to vary, for example by treatment group. The *intracluster correlation* (*ICC*) is the ratio of the between-cluster variance to the total variance of an outcome variable and is often denoted by ρ. From model (1.1) we can see the total variance (V) is given by $\sigma_\alpha^2 + \sigma_\epsilon^2$.

Thus,

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2} \tag{1.2}$$

The ICC quantifies the correlation between the outcomes of any two individuals within the same cluster. It should be noted that, like an ordinary correlation coefficient, ρ cannot exceed unity. This happens when $\sigma_\epsilon^2 = 0$, in other words when all the values in a cluster are the same, that is all individuals in the same cluster are identical with respect to the outcome, then $\rho = 1$. If the values in each cluster are the same, then $\rho = 0$; that is individuals within the same cluster are no more likely to have similar outcomes than individuals in different clusters. This might happen if in fact the outcome was at the cluster level. However, unlike an ordinary correlation coefficient ρ cannot go negative and is bounded by zero, when $\sigma_\alpha^2 = 0$, which occurs when all the clusters have the same mean value. However, for real data, the ICC can be negative. This can occur when there are limited resources to share within clusters, and so if one person in a cluster gets a large share of resources, then another in the same cluster will get a smaller share of resources. For example, if a sow is feeding a large number of piglets, then there is often a runt who gets less than its fair share. This may also happen when doctors treat a number of patients, more the time spent with one patient is less time for another. In this case, model (1.1) may not be realistic.

Since σ_α^2 is the covariance between any two observations within a cluster, in the case of when there are only two observations per cluster, the ICC can be estimated by the Pearson's correlation coefficient.

1.10.3 Estimating the intracluster correlation (ICC) coefficient

There are a number of ways of estimating the ICC. The simplest, which works best with balanced data, is from an analysis of variance (ANOVA) model and is described in Appendix 1.A.1. One can also derive the ICC from a random-effects model. We will describe this in more detail in Chapter 5, but in the meantime we give the R code to do it in Appendix 1.A.3. There is also a publically available R package 'ICC' which estimates the ICC and gives a confidence interval about the estimate which is also described in Appendix 1.A.3.

As an example consider a random set of data generated according to model (1.1). The R program *genranddata* (given in Appendix 1.A.3) generates a set of random variables according to model (1.1) with $\beta_0 = \beta_1 = 0$ and $\sigma_\alpha = 1$ and $\sigma_\epsilon = 4.36$ so that $\rho = 1^2/(1^2 + 4.36^2) = 0.05$. The R program *calcicc* (also in Appendix 1.A.3) uses both methods and gives the results in Textbox 1.1.

We can see that the estimates from both methods are very close and also quite close to the population ICC of $\rho = 0.05$. One should not expect exactly the same results since the methods are different. However, most methods give similar results unless the data are very extreme (outliers and unbalanced cluster sizes), in which case all methods are challenged. We consider the binary case in the next section.

Textbox 1.1

Output from R program *calcicc* using data generated by *genranddata*

ICC derived from one-way ANOVA	0.04645379
ICC derived from random-effects model	0.04645367

1.10.4 Link between the Pearson correlation coefficient and the intraclass correlation coefficient

Most people are familiar with the Pearson correlation coefficient and it is worth investigating the link between the two. Suppose we have two random variables X_1 and X_2 which we think of these as being sampled from the same cluster i. Suppose $E(X_1) = E(X_2) = 0$ and $\text{Var}(X_1) = \text{Var}(X_2) = \sigma^2$, $E(X_1X_2) = \sigma^2\rho_{\text{Pearson}}$ where ρ_{Pearson} is the Pearson correlation coefficient.

Using model (1.1) we can write X_1 as $a_i + e_{i1}$ and X_2 as $\alpha_i + \epsilon_{i2}$ and so $\sigma^2 = \sigma_\alpha^2 + \sigma_\epsilon^2$ Then

$$E(X_1 - X_2)^2 = E(\epsilon_{i1} - \epsilon_{i2})^2 = 2\sigma_\epsilon^2$$

but

$$E(X_1 - X_2)^2 = E({X_1}^2) - 2E(X_1X_2) + E({X_2}^2) = 2\sigma^2 - 2\sigma^2\rho_{\text{Pearson}}$$

and so

$$\rho_{\text{Pearson}} = \frac{(\sigma^2 - \sigma_\epsilon^2)}{\sigma^2} = \frac{\sigma_\alpha^2}{\sigma^2} = \text{ICC}$$

Thus under model (1.1) the ICC and the Pearson correlation coefficient are the same. It is perhaps worth remarking that before Fisher derived the ANOVA, the intraclass correlations were used as an extension of Pearson's correlations to investigate comparisons of more than two groups. However, there is a fundamental difference between the ICC and ρ_{Pearson} in that the Pearson's correlation is based on a regression model $y = a + bx$. It does not matter whether we substitute X_1 for y and X_2 for x, or vice versa, we get the same correlation, but the regression model allows b to be negative. In model (1.1), the observations are also 'exchangeable' in that it does not matter which one is written first, and if one exchanged the first for the second in any cluster it would make no difference to the intraclass correlation. However model (1.1) does not allow negative correlations which are permissible under a linear regression model. In addition,

X_1 and X_2 have the same expectation which is not true for a regression model. A technique to obtain an estimate which might be expected to give a closer value to the population value would be the following. If $x1$ and $x2$ are the first and second vectors of observations of length n, concatenate $x1$ and $x2$ to give a new vector $x12$ of length $2n$ and also concatenate $x2$ and $x1$ to give a new vector $x21$. This will also have expected value ρ, but now the mean value of the two vectors is the same. We have written an R program *pearsonicc* to find the Pearson's correlation using the usual methods, and the concatenated method and the results are given in Textbox 1.2. One can see that these estimates are close and give a reasonable estimate of the population ICC.

Textbox 1.2

Output from R program *pearsonicc*

Correlation coefficient between $x1$ and $x2$	0.04298662
Correlation between $x1:x2$ and $x2:x1$	0.04143834

If the ICC is small or in fact if $\sigma_\alpha^2 = 0$, then by chance an estimate may be negative. If the data were in fact generated by model (1.1), then one cannot have $\rho = 0$, but in practice of course, one does not know which model generated the data, and negative ICCs can be observed. Examples of this are when there are finite resources in a cluster, so that if one individual gets more resource then another gets less. For example, the weights of piglets in a litter may be negatively correlated, or there may be a finite amount of money for medical care in a health centre, so that if a large amount of money is spent on one individual, much less has to be spent on others. A review of the numerical values of ICCs likely to be seen in primary care has been given by Adams *et al*. (2004). In general, values between 0.01 and 0.05 are found, but larger values can occur if the outcome is in some way associated with the cluster. For example, some doctors are more likely to issue a prescription for antibiotics for a sore throat than others and so if the outcome was an antibiotic prescription, one might expect quite a high ICC.

1.11 Random-effects models for binary outcomes

1.11.1 The model

For a binary outcome, we have a dependent variable y_{ij} with expectation π_{ij} for the ith patient in the jth cluster. The basic random-effects logistic model is

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 x_{ij} + \alpha_j \tag{1.3}$$

Logistic regression is discussed in many books including Campbell (2006). Conventionally, we assume that α_j is Normally distributed with variance σ_α^2.

An alternative is to assume that the π_{ij} are distributed with a Gamma distribution and conditional on π_{ij} and we assume the counts have a Poisson distribution. In this case, the observed counts will have a negative binomial distribution.

1.11.2 The ICC for binary data

With no individual covariates, we assume that each cluster j has a probability of success of π_j. There are a number of possibilities for the ICC. The simplest is to assume that the π_j have a constant variance σ_α^2. We also assume that a randomly selected individual from the population has a probability of success of π, where π is the probability of success for the whole population. If we assume that the estimate for π has variance $\pi(1-\pi)$, then the intracluster correlation is defined as

$$\rho = \frac{\sigma_\alpha^2}{\pi(1-\pi)} \tag{1.4}$$

Another formulation is to assume a model of the form Equation 1.3. In this case, $\text{var}(\log(\pi/(1-\pi))) = 1/(\pi(1-\pi)$ and so for the logistic model $\rho = \pi(1-\pi)\sigma_\alpha^2$. It is important to note the different formulation depending on different models.

As with the case for continuous data, we can estimate the ICC in a number of ways. Perhaps surprisingly, carrying out an ANOVA in which the dependent variable is simply 0 or 1 gives a valid estimate of ρ, despite the usual Normality assumptions not being met. However, it is can be estimated using a random-effects model of the form Equation 1.3 in a similar fashion to that described for the linear model in Section 1.10.2. This is described in more detail in Chapter 6.

Unlike the continuous case, it can be seen that for binary data the ICC will depend on the proportion π, which might be an incidence or a prevalence. If the model (1.4) holds true and if π is very small, then so will be the ICC. Gulliford *et al*. (2005) looked at empirical evidence of the result and gave examples of ICCs likely to occur with chosen binary variables in practice. From 188 ICCs for binary outcomes in general practice, they found the median prevalence to be 13.1% (interquartile range IQR 3.5–28.4%) and median ICC 0.051 (IQR 0.011–0.094). There were 136 ICCs from a Health Technology Assessment (HTA) review, with median prevalence 6.5% (IQR 0.4–20.7%) and median ICC 0.006 (IQR 0.0003–0.036). There was an approximate linear association of log ICC with log prevalence in both data sets. When the prevalence was 1%, the predicted ICC was 0.008 from the general practice or 0.002 from the HTA, but when the prevalence was 40% the predicted ICC was 0.075 (GPRD, general practitioner research database) or 0.046 (HTA). They concluded that the prevalence of an outcome may be used to make an informed assumption about the magnitude of the intraclass correlation coefficient.

1.11.3 The coefficient of variation

The two sources of variation are captured by σ_α^2 and σ_ϵ^2. The problem is that these will vary depending on the units of measurement. One advantage of ρ is that it is dimensionless and can be assumed similar for different outcome measures.

Hayes and Moulton (2009) define a different dimensionless variable to summarize the variability between clusters: the coefficient of variation of clusters (cvc). This is described as the standard deviation between clusters divided by the overall mean of the outcome. Using the notation of Equation 1.1 and assuming no covariates, we have

$$\text{cvc} = \frac{\sigma_\alpha}{\beta_0} \tag{1.5}$$

One advantage of the cvc is that it is applicable for continuous variables, rates and proportions. For proportions, we replace μ with π where π is the overall proportion of successes:

$$\text{cvc} = \frac{\sigma_\alpha}{\pi} \tag{1.6}$$

1.11.4 Relationship between cvc and ρ for binary data

From Equations 1.3 and 1.6, we get that

$$\rho = \frac{\text{cvc}^2 \pi}{1 - \pi}$$

Since $\rho \leq 1$, this also means that

$$\text{cvc} \leq \sqrt{\frac{1 - \pi}{\pi}}$$

This can be useful if one has a good idea about the value of π, but are rather vague about the value of the cvc.

Thomson, Hayes and Cousens (2009) argue that practical experience suggests that epidemiologists may find the cvc easier to use and interpret, since the coefficient of variation (the standard deviation expressed relative to the mean) is a familiar concept used in many branches of medicine, whereas few find the intracluster correlation coefficient as an intuitive measure. When the outcome is continuous, ρ is the more appropriate measure of between-cluster variability. Knowledge of ρ is sufficient to correct for clustering in a sample size calculation, (Chapter 3), but knowledge of cvc is not sufficient on its own and a measure of within-cluster variability is also necessary. Conversely, when the outcome is a person-time rate, cvc is readily defined, but ρ is not as there is no clearly defined unit of observation when working with a person-time denominator. In Chapter 6, we will discuss how we can use different outcome measures for binary data, such as an odds ratio or a relative risk. If we assume that an intervention has a constant effect on one such measure, then we cannot necessarily assume the measure of intracluster variation is constant. Thomson *et al*. (2009) show that in some circumstances, it is more likely that cvc will be constant than ρ.

We suggest that, in general, one should use ρ, but be aware that in what we have termed field trials, investigators may find cvc more useful.

1.12 The design effect

The *design effect* (*DE*) is the ratio of the variance of an outcome measure when clustering is accounted for to the variance of the outcome measure when clustering is not accounted for. It is often referred to as the *variance inflation factor* (*VIF*) since it measures the amount that one should increase a variance estimate obtained by ignoring clustering to allow for the clustering effect. For clusters of equal size m, it can be shown that $\text{DE} = 1 + (m - 1)\rho$. The derivation is shown in Appendix 1.A.2. This is also called the *sample size inflation factor (SSIF)* since the same factor that inflates the variance will also inflate the required sample size. This is because the required sample size for a trial is directly proportional to the variance of the outcome measure.

Suppose we calculated the sample size for an individually randomised trial (using the effect size, power and significance level) and found we needed n patients per arm. For a cluster randomised trial with the same parameters, we would need $n \times \text{DE}$ patients per arm.

It is commonly believed that since the ICC is likely to be small, then the DE is also going to be small. However, this is to ignore the fact that the DE also includes the number of patients per cluster. If we only have one patient per cluster, that is $m = 1$, we have in effect an individually randomised trial and the $\text{DE} = 1$. However, suppose $\rho = 0.05$, which is by no means uncommon in trials in primary care. If we have $m = 21$, that is 21 subjects per cluster, then we find the $\text{DE} = 2$. This means that a cluster trial with this DE would have to recruit *twice* as many patients as the corresponding individually randomised trial. Correspondingly, if we analysed a cluster trial but conducted an analysis appropriate for an individually randomised trial, the standard error of our parameter estimates would be underestimated by $\sqrt{2}$, and so we would be more likely to declare factors statistically significant than expected from the preset Type I error.

The consequence of this is that one needs to weigh up carefully the counterbalancing needs of unbiased estimates of the treatment effect, which a cluster trial is more likely to deliver versus improved precision, which an individually randomised trial will give. In general, since bias is usually unknown, one would prefer to have an estimate of the treatment effect that is believed to be unbiased, albeit slightly less precise, and so one would choose a cRCT design in preference to an individually randomised design. Puffer *et al*. (2003) in particular argue for more careful consideration of the options before adopting a cluster trial.

1.13 Commonly asked questions

Should I always use a cluster trial to evaluate an intervention, if both a cluster and individually randomised trial are possible?

The general answer is if you can be sure to avoid contamination and can deliver the intervention to an individual, then it may be possible, and indeed better, to use an individually randomised trial. An example of an individually randomised trial that might have been clustered is the HOOP trial (McCandlish *et al*., 1998). The aim of this trial was to compare the effect of two methods of perineal management used during spontaneous vaginal delivery on the prevalence of perineal pain reported at 10 days after birth. At the end of the second stage of labour, women were allocated to either the 'hands on' method, in which the midwife's hands put pressure on the baby's head and support ('guard') the perineum; lateral flexion is then used to facilitate delivery of the shoulders or the 'hands poised' method, in which the midwife keeps her hands poised, not touching the head or perineum, allowing spontaneous delivery of the shoulders. The midwives' compliance with the allocated policies of hands off or hands poised was reported as being around 70%, and the likelihood of contamination between the treatment groups is high.

1.14 Websources

A video of Allan Donner talking about the basics of cluster trials is given here: http://www.youtube.com/watch?v=I4gfGl_JBEA

There are also videos of other aspects of cluster trials by Sandra Eldridge, Mike Campbell, Ian White and Charles Weijer: http://www.newton.ac.uk/programmes/DAE/seminars/.

The discussion of ethics in cluster trials can be found at: http://crtethics.wikispaces.com.

Exercise

Answer the following as true of false

1. In a cluster randomised trial, the main purposes for adopting a cluster design may be to:

(a) Minimize contamination

(b) Get more power than from an individually randomised trial of same size

(c) Blind patients to the intervention

(d) Effective delivery of the intervention.

2. Field trials

(a) Usually have more clusters than cohort trials

(b) Can be known as *cluster–cluster trials*

(c) Follow the same subjects over time

(d) May be more representative of a population than a cohort trial.

Appendix 1.A

1.A.1 Estimating the ICC from an analysis of variance

Let MSB and MSW denote the mean square between and mean square within clusters from a standard one-way ANOVA. Then the usual ANOVA table for Equation 1.1 is as follows:

Source	Sum of squares	Degrees of freedom	Mean square	Estimates
Between clusters	SS_B	$k-1$	MSB	$m\sigma_\alpha^2+\sigma_\epsilon^2$
Within clusters	SS_W	$k(m-1)$	MSW	σ_ϵ^2
Total		$mk-1$		

Then we have

$$\widehat{\rho}=\frac{\text{MSB}-\text{MSW}}{\text{MSB}+(m-1)\text{MSW}}=\frac{S_B^2}{S_B^2+S_W^2}$$

where

$$S_W^2=\text{MSW}\quad S_B^2=\frac{\text{MSB}-\text{MSW}}{m}$$

are the sample estimates of σ_ϵ^2 and σ_α^2, respectively.

1.A.2 Derivation of the design effect

Suppose in model (1.1), there were no covariates and all the clusters are of the same size, so the simple model is

$$y_{ij} = \alpha_j + \epsilon_{ij} \tag{A1}$$

where α_j and ϵ_{ij} are assumed independent random variables and where $j = 1, \ldots, K$ are the index clusters and $i = 1, \ldots, m$ are the index subjects. Then $V(y_{ij}) = \sigma^2 = \sigma_\alpha^2 + \sigma_\epsilon^2$. The presence of α_j means that y_{ij} and $y_{i'j}$ are correlated when $i \neq i'$, and since $E(\alpha_j) = E(\epsilon_{ij}) = 0$ we have $\text{Cov}(y_{ij}, y_{i'j}) = E(y_{ij}y_{i'j}) = \sigma_\alpha^2$.

Now consider

$$\text{Var}(\bar{y}) = \text{Var}\left(\frac{\left(\sum_{j=1}^{K}\sum_{i=1}^{m} y_{ij}\right)}{mk}\right) = \frac{1}{(mk)^2}\sum_{j=1}^{K}\text{Var}\left(\sum_{i=1}^{m} y_{ij}\right) \tag{A2}$$

Since the first summation is over clusters and the clusters are independent, we can simply sum over the clusters and take the variance inside the summation.

Now

$$\text{Var}\left(\sum_{i=1}^{m} y_{ij}\right) = E\left\{\left(\sum_{i=1}^{m} y_{ij}\right)^2\right\} = m^2\sigma_\alpha^2 + m\sigma_\epsilon^2$$

and so Equation A2 becomes

$$\text{Var}(\bar{y}) = \frac{m^2 k\sigma_\alpha^2 + mk\sigma_\epsilon^2}{(mk)^2} = \left(\frac{mk\left(m\sigma_\alpha^2 + \sigma_\epsilon^2\right)}{(mk)^2} = \frac{\sigma^2}{mk}(1 + \rho(m-1)\right)$$

where $\rho = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\epsilon^2)$. The term σ^2/mk is the variance of the mean under independence and so the $(1 + \rho(m-1))$ is the amount we need to inflate the variance under model (1.1).

1.A.3 R programs

Further explanation of the implementation of R is given in Chapter 10.

1.A.3.1 R code to generate Figure 1.1

Program stripchart

```
setwd("f:\\Cluster RCTs book\\Data\\R data sets")
desmond<-read.table("desmond.txt", header=T)
attach(desmond)
names(desmond)

hba1c1n<-as.numeric(as.character(hba1c1))

par(mfrow=c(2,1))
```

```
stripchart(hba1c1n[random=="I"]~pract[random=="I"],pch=21,
main="Intervention Practices k=24, patients n=296",
xlab="HbA1c at 1 year", ylab="Practice number")
stripchart(hba1c1n[random=="C"]~pract[random=="C"],pch=21,
main="Control Practices k=24, patients n=340", xlab="HbA1c
at 1 year", ylab="Practice number")
```

1.A.3.2 Program to generate clustered data

Program *genranddata*

This program generates $k = 100$ clusters of size $m = 2$, with an ICC of 0.05.

```
setwd("f:\\Cluster RCTs book\\Data\\R data sets")


#generate 100 clusters each of size 2
k<-100
m<-2

# generate the clusters
c<-gl(k,m)
n<-m*k

#generate the cluster random effects
a1<-rnorm(k,0,1)
a11<-rep(a1, each=m)
#generate the individual random noise
e1<-rnorm(n,0,4.36)
#Now generate the sum
x<-a11+e1

data<-data.frame(x,c)
write.table(data, "rand.txt", col.names=c("x","c"),row.names=F)
```

1.A.3.3 Program to calculate the ICC from raw data

Program *calcicc*

This program estimates the ICC generated by *genranddata* using ANOVA and linear mixed models.

```
setwd("f:\\Cluster RCTs book\\Data\\R data sets")
random<-read.table("rand.txt", header=T)
attach(random)
names(random)

# One analysis of variance between clusters

aovmod<-aov(x~c)

#extract the mean squares
ms<-summary(aovmod)[[1]][[3]]
```

```
#find variance components
s2w<-ms[2]
s2b<-(ms[1]-ms[2])/m
rho<-s2b/(s2w+s2b)

names(rho)<-"ICC derived from one way anova"
rho
#Using a linear mixed model we need to specify that the constant
is fixed and clusters are random
library(nlme)

lmemod<-lme(fixed =x~1, random=~1|c)
s2b1<-as.numeric(VarCorr(lmemod))[[1]]
s2w1<-as.numeric(VarCorr(lmemod))[[2]]
rho1<-s2b1/(s2b1+s2w1)
names(rho1)<- "ICC derived from random effects model"
rho1
```

1.A.3.4 Program to calculate the compute the ICC for paired data using Pearson's correlation

Program *pearsonicc*

This program estimates the ICC from the data generated by genranddata using two forms of the Pearson correlation.

```
setwd("f:\\Cluster RCTs book\\Data\\R data sets")
random<-read.table("rand.txt", header=T)
attach(random)



# Finding pearson's correlation between the first and second vari-
ables
# Generate a variable which indicates the first and second observa-
tion in a cluster

var<-gl(2,1,200)
x1<-x[var==1]
x2<-x[var==2]
r<-cor(x1,x2)
names(r)<- "correlation coefficient between x1 and x2 "
r

# It is better to ensure the means are the same for the two vari-
ables.
# This can be achieved by concatenating x1 and x2 and also x2
with x1 and correlating the extended vectors
```

```
x12<-c(x1,x2)
x21<-c(x2,x1)
r1<-cor(x12,x21)
names(r1)<-" correlation between x1:x2 and x2:x1"
r1
```

1.A.3.5 Use of user-supplied program ICC

A package 'ICC' by Matthew Wolak including a number of functions to assist in the estimation of the ICC is available in R. Functions included an estimate of the ICC from variance components of a one-way ANOVA and also estimate the number of individuals or groups necessary to obtain an ICC estimate with a desired confidence interval width.

One needs to 'load' and 'install' ICC. One of the packages is ICCest which estimates the ICC and gives a 95% confidence interval and the results are shown below.

```
> ICCest(c,x, data, 0.05, "THD")
$ICC
[1] 0.04652074
$LowerCI
[1] -0.1498308
$UpperCI
[1] 0.2395293
$N
[1] 100
$k
[1] 2
$varw
[1] 20.8661
$vara
[1] 1.018068
```

The confidence intervals using the method 'THD' (Thomas, Hultquist and Donner) are not ideal since they allow negative numbers, especially when the ICC is low and the numbers of subjects per cluster is small. In this case, the second method from calcicc is better.

Using the same data, our program calcicc gives the same result for the estimate.

```
> rho
ICC derived from one way anova
0.04652074
```